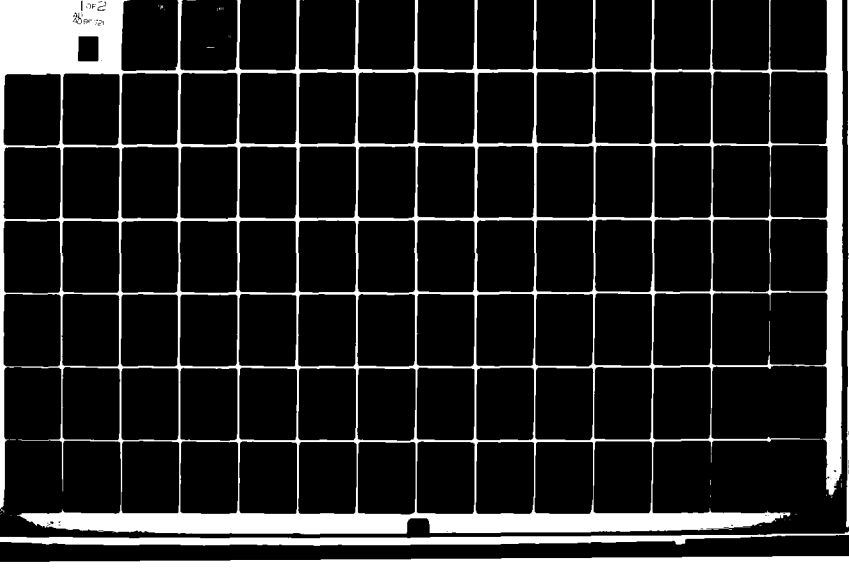
AD-A095 721 TECHNOLOGY SERVICE CORP SANTA MONICA CA F/G 12/1
FURTHER DEVELOPMENT OF NEW METHODS FOR ESTIMATING TAIL PROBABIL--ETC(U)
JAN 81 L BREIMAN, C J STONE, J D GINS F49620-80-C-0037
UNCLASSIFIED TSC-PD-A243-1 AFOSR-TR-81-0136 NL

1 OF 2
AD A095 721



AFOSR-TR- 81-0136

TW

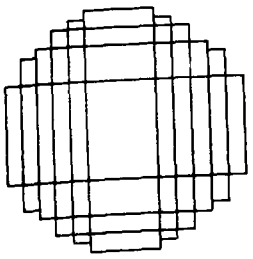
(12)

LEVEL II

AD A095871

DTIC FILE COPY

DTIC
ELECTE
MAR 0 2 1981
S D
E



Technology Service Corporation

distribution unlimited.

TW

81 2 27 078

(12) AFOSR
(16) 2241
(17) 81-0136
(17) A5

LEVEL II (12)

(9) Final Rept.

FURTHER DEVELOPMENT OF NEW METHODS FOR ESTIMATING TAIL PROBABILITIES AND EXTREME VALUE DISTRIBUTIONS.

(14) TSC-PD-A243-1
January 1981
(11)
(12) 119

L. Breiman
C. J. Stone
J. D. Gins

FINAL REPORT

Contract [redacted] F49620-80-C-0037

Submitted to:

U.S. Air Force
Office of Scientific Research
Bolling Air Force Base
Building 410
Washington, D.C. 20332

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) UNCLASSIFIED

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 81-0133	2. GOVT ACCESSION NO: AD-A095 721
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): FURTHER DEVELOPMENT OF NEW METHODS FOR ESTIMATING TAIL PROBABILITIES AND EXTREME VALUE DISTRIBUTIONS	5. TYPE OF REPORT & PERIOD COVERED: Final
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): L. Breiman C. J. Stone J. D. Gins	8. CONTRACT OR GRANT NUMBER(s): F49620-80-C-0037
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Technology Service Corporation, 2950 Thirty-first Street, Santa Monica, Los Angeles County, CA 90405	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: January 1981
	13. NUMBER OF PAGES: 109
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

extreme quantile
confidence interval
exponential tail
tail heaviness
Monte Carlo

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This research has focused on the problem of obtaining confidence intervals for extreme quantiles based on a random sample form a distribution of unknown form. Three confidence interval procedures were studied, both analytically and by means of an extensive Monte Carlo experiment. The experiment involved three sampled sizes (100, 200, 400) and twenty underlying distributions (five Weibulls, five mixed Weibulls, five lognormals, five mixed lognormals). The Monte Carlo results show that all three procedures studied work quite well and point the way to further improvement.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) UNCLASSIFIED

CONTENTS

Section

Appendix

1. INTRODUCTION

Given a positive number $q<<1$ let x_q denote the (upper) q^{th} quantile of a random variable X, defined by $P(X \geq x_q)=q$. In a previous report (Breiman, Stone and Gins [2]) a number of point estimators of x_q based on a random sample of size n from the distribution of X were studied. The main purpose of the present research is to develop and study the performance of several confidence interval procedures for x_q based on the sample, especially when $q=1/n$. This work is reported on in Section 2. In Subsection 2.1.1, modifications of one of these procedures are developed for handling grouped data, which contain many ties, satisfactorily. A detailed discussion of the quadratic tail procedure, described in Subsection 2.3, is given in Section 3. In Section 4 the behavior of the exponential tail estimator is studied when the sample data is stationary but dependent. The selection of procedures for consideration has been guided by the dictum of DuMouchel and Olshen [3] that one should "let the tails of the data speak for themselves."

2. CONFIDENCE INTERVALS FOR EXTREME QUANTILES

Let X denote a random variable whose (unknown) underlying distribution function F is continuous on $(-\infty,\infty)$ and strictly increasing on an interval I which contains the support of F [i.e., which is such that $P(X\varepsilon I)=1$]. Let $S(x)$, $x\geq 0$, be the tail (survival) probability function defined by $S(x)=P(X>x)=1-F(x)$. For $0<q<1$ let x_q denote the (upper) q^{th} quantile of X defined by

$$S(x_q) = 1-F(x_q) = P(X>x_q) = q \quad .$$

Let $X_1,\ldots,X_n$ denote a random sample of size n from F and let $X_{(1)},\ldots,X_{(n)}$ denote the corresponding order statistics defined so that $X_{(1)}>\ldots>X_{(n)}$. A confidence interval procedure for x_q is an interval $\bar{I} = [\underline{x}_q,\bar{x}_q]$, where $\underline{x}_q<\bar{x}_q$ with both $\underline{x}_q$ and $\bar{x}_q$ being functions of $X_{(1)},\ldots,X_{(n)}$. The length of $|I|$ of the interval I is given by $I = \bar{x}_q-\underline{x}_q$.

Let P_F and E_F denote probabilities and expectations when F is the distribution function of X. Relevant characteristics of the confidence interval procedure include $P_F(x_q<\bar{I})$, $P_F(x_q>\bar{I})$, $P_F(x_q\notin\bar{I}) = P_F(x_q<\bar{I}) + P_F(x_q>\bar{I})$, and $E_F|I|$. (Here $x_q<\bar{I}$ and $x_q>\bar{I}$ mean respectively that $x_q<\underline{x}_q$ and $x_q>\bar{x}_q$.) The goals of making the indicated probabilities and expected length both small over a wide range of F's are obviously in conflict with each other.

One approach to obtaining a confidence interval procedure is to start out with a realistic model $F(\cdot;\theta)$, $\theta\varepsilon\Theta$, for F, express x_q as a function $x_q(\theta)$ of the unknown parameter θ, and then employ a classical parametric confidence interval procedure for $x_q(\theta)$. Given $0<\alpha<1$, suppose $\bar{I}$ is (at least approximately) a classical $100(1-\alpha)\%$ confidence interval based on the assumed model; i.e., that

$$P_\theta(x_q \notin \bar{I}) \doteq \alpha, \qquad \theta \in \textcircled{H},$$

where $P_\theta = P_{F(\cdot;\theta)}$. Then roughly speaking, $P_F(x_q \notin \bar{I}) \doteq \alpha$ for any F which can be globally well approximated by $F(\cdot;\theta)$ for some $\theta \in \textcircled{H}$. Otherwise $P_F(x_q \notin \bar{I})$ can be substantially larger than α. This is true in particular for F's whose central portion and extreme upper tail are best approximated by distribution functions $F(\cdot;\theta)$ with significantly different values of θ. Relative to such F's the confidence interval procedure is not robust vis a vis

$$P_F(x_q \notin \bar{I}) \doteq \alpha \qquad . \tag{2.1}$$

There are two ways of making such a confidence procedure more robust: (1) Consider a higher dimensional model $F(\cdot;\theta,\tau)$, $(\theta,\tau) \in \textcircled{H} \times T$ (e.g., a three-parameter instead of a two-parameter lognormal model). A wider range of distribution functions F can be globally well approximated by a distribution function in this larger model. (2) Base the confidence interval procedure on only the upper m order statistics $X_{(1)},\ldots,X_{(m)}$ for some $m<n$, rather than on all the original data. Both (1) and (2) lead to procedures which are more robust vis a vis Eq. (2.1). Unfortunately they also both lead to significant increases in $E_F \bar{I}$.

Robustness vis a vis Eq. (2.1) here is with respect to departures of F from the assumed model. Another type of robustness, when F belongs to the assumed model, is with respect to errors in measuring or recording the sample data; this type of robustness will not be considered in the present report.

[Note that (1) and (2) above can lead to procedures which are less (not more!) robust with respect to measurement and recording errors.]

Understandability and ease of implementation are important additional considerations in determining which procedures to use.

So far, confidence intervals $\overline{I}=\overline{I}(q)$ have been considered for quantiles x_q, $0<q<1$. It is also desirable to consider confidence intervals $\overline{J}=\overline{J}(x)$ for tail probabilities $S(x)=1-F(x)$, $-\infty<x<\infty$. There is a natural one-to-one correspondence between confidence intervals for x_q and these for $1-F(x)$ given by $\overline{J}(x)=\{q:x\epsilon\overline{I}(q)\}$ and $\overline{I}(q)=\{x:q\epsilon\overline{J}(x)\}$. This correspondence preserves coverage probabilities. That is,

$$P_F\left(x_q\epsilon\overline{I}(q)\right) = P_F\left(q\epsilon\overline{J}(x_q)\right)$$

and

$$P_F\left(1-F(x)\epsilon\overline{J}(x)\right) = P_F\left(x_{1-F(x)}\epsilon\overline{I}\left(1-F(x)\right)\right) \text{ if } 0<F(x)<1 \quad .$$

Because of this close correspondence between confidence intervals for quantiles and those for tail probabilities it was decided to devote the present research effort exclusively to confidence intervals for quantiles.

Three confidence interval procedures will be described in Subsections 2.1 through 2.3. (The procedure described in Subsection 2.3 will be elaborated on in Section 3.) A Monte Carlo experiment designed to compare the performance characteristics of these procedures will be discussed in Subsection 2.4. Tentative conclusions drawn from the results of this experiment and suggestions for further work are presented in Subsection 2.5.

2.1 TWO-PARAMETER EXPONENTIAL PROCEDURE

Consider the two-parameter exponential model

$$F(x;\tau,a) = 1-e^{-(x-\tau)/a}, \quad x\geq\tau,$$

$$= 0, \quad x<\tau,$$

where $\tau\varepsilon(-\infty,\infty)$ is a location parameter and $a>0$ is a scale parameter. Correspondingly

$$S(x;\tau,a) = e^{-(x-\tau)/a}, \quad x\geq\tau,$$

$$= 1, \quad x<\tau,$$

and

$$x_q = \tau+a \log(1/q), \quad 0<q<1 \quad .$$

The two-parameter exponential model is appealing for several reasons. First, it is very simple and leads to the above simple formula for x_q. Second, the upper tail of distributions of this type can be used to provide reasonably accurate approximations to the upper tail of a number of commonly assumed alternative models--Weibull, gamma, and lognormal. Third, the upper tail of distributions of this type is realistic in many applications (see, e.g., Breiman, Gins, and Stone [1]).

Given $2 \leq m \leq n$, let $X_{(1)},\ldots,X_{(m)}$ denote the upper m order statistics based on a random sample of size n from $F(\cdot;\tau,a)$. The joint density of these random variables is given by

$$f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;\tau,\mu)$$

$$= \frac{n!}{(n-m)!}\, a^{-m} e^{-\sum_1^m (x_i - x_m)/a}\, e^{-m(x_m-\tau)/a} \left[1-e^{-(x_m-\tau)/a}\right]^{n-m}$$

for $\tau < x_m < \ldots < x_1$, while $f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;\tau,a)=0$ otherwise. The maximum likelihood estimators of τ and a based on $X_{(1)},\ldots,X_{(m)}$ are given by

$$\bar{a} = \frac{1}{m} \sum_1^{m-1} [X_{(i)} - X_{(m)}]$$

and

$$\bar{\tau} = X_{(m)} - \bar{a} \log \frac{n}{m} \quad .$$

The corresponding maximum likelihood estimator of x_q is given by

$$\bar{x}_q = \bar{\tau} + \bar{a} \log \frac{1}{q} = X_{(m)} + \bar{a} \log \frac{m}{nq} \quad .$$

This estimator, with $\bar{a}$ replaced by

$$\hat{a} = \frac{1}{m-1} \sum_1^{m-1} [X_{(i)} - X_{(m)}] \quad ,$$

was studied in detail in Breiman, Stone, and Gins [2], where it was referred to as the exponential tail estimator. Considering its simplicity, it is surprisingly robust to moderate departures of the upper tail of F from the assumed exponential form.

A confidence interval procedure for x_q will now be obtained. To this end set $Y_i=X_{(i)}-X_{(m)}$ for $1\le i\le m-1$. Then

$$f_{Y_1,\ldots,Y_{m-1},X_{(m)}}(y_1,\ldots,y_{m-1},x_m;\tau,a)$$

$$= \frac{n!}{(n-m)!}\, a^{-m}\, e^{-\sum_1^{m-1} y_i/a}\, e^{-m(x_m-\tau)/a}\left[1-e^{-(x_m-\tau)/a}\right]^{n-m}$$

for $x_m>\tau$ and $0<y_{m-1}<\ldots<y_1$, while $f_{Y_1,\ldots,Y_{m-1},X_{(m)}}(y_1,\ldots,y_{m-1},x_m;\tau,a)=0$ otherwise. Let $Z_1,\ldots,Z_{m-1}$ be independent random variables having the common exponential density f_{Z_1} defined by

$$f_{Z_1}(z) = \frac{1}{a}\, e^{-z/a}\,, \qquad z>0\,,$$

$$= 0\,, \qquad z\le 0\,,$$

and suppose that $Z_1,\ldots,Z_{m-1},X_{(m)}$ are independent random variables. Let $Z_{(1)},\ldots,Z_{(m-1)}$ be the order statistics from $Z_1,\ldots,Z_{m-1}$, defined so that $Z_{(1)}>\ldots>Z_{(m-1)}$. Then $Z_{(1)},\ldots,Z_{(m-1)},X_{(m)}$ has the same joint distribution

as $Y_1,\ldots,Y_{m-1},X_{(m)}$. Consequently $Y_1+\ldots+Y_{m-1}$ is independent of $X_{(m)}$ and has the same distribution as

$$Z_{(1)}+\ldots+Z_{(m-1)} = Z_1+\ldots+Z_{m-1} \quad ,$$

namely the gamma distribution having density $g(\cdot;m-1,1/a)$ defined by

$$g(t;m-1,1/a) = \frac{t^{m-2}e^{-t/a}}{a^{m-1}(m-1)!} , \quad t>0 \quad ,$$

$$= 0 , \quad t\leq 0 \quad .$$

Therefore $\hat{a}$ has the gamma distribution with density $g\left(\cdot;(m-1),(m-1)/a\right)$. The density of $X_{(m)}$ is given by

$$f_{X_{(m)}}(x;\tau,a) = \frac{n!}{(m-1)!(n-m)!}\frac{1}{a}e^{-m(x-\tau)/a}\left[1-e^{-(x-\tau)/a}\right]^{n-m}$$

for $x_m>\tau$ and $f_{X_{(m)}}(x;\tau,a)=0$ for $x\leq\tau$. The distribution function of $X_{(m)}$ is given by

$$F_{X_{(m)}}(x;\tau,a) = 1-B\left(e^{-(x-\tau)/a};m,n-m+1\right) \quad ,$$

where

$$B(p;m,n-m+1) = \int_0^p \frac{n!}{(m-1)!(n-m)!} r^{m-1}(1-r)^{n-m}\,dr \quad .$$

Let $-\infty<M<\infty$. Since $X_{(m)}$ and $\hat{a}$ are independent, for $-\infty<z<\infty$

$$P_{\tau,a}\left(\tau+Ma\leq X_{(m)}+z\,\hat{a}\right)$$

$$=\int_0^\infty P_{\tau,a}\left(X_{(m)}\geq\tau+Ma-zt\right)f_{\hat{a}}(t;\tau,a)\,dt$$

$$=\int_0^\infty B(e^{-M}e^{zt};m,n-m+1)g(t;m-1,m-1)\,dt \quad .$$

Now $x_q = \tau + a\log(1/q)$, so that

$$P_{\tau,a}\left(x_q\leq X_{(m)}+z\,\hat{a}\right)=\int_0^\infty B(qe^{zt};m,n-m+1)g(t;m-1,m-1)\,dt \quad .$$

Given $0<\alpha<1$, define $z_\alpha = z_\alpha(q;m,n)$ by

$$\int_0^\infty B(qe^{z_\alpha t};m,n-m+1)g(t;m-1,m-1)\,dt=\alpha \quad . \tag{2.2}$$

Then

$$P_{\tau,\mu}\left(x_q\leq X_{(m)}+z_\alpha\,\hat{a}\right)=\alpha \text{ for } -\infty<\tau<\infty \text{ and } a>0 \quad . \tag{2.3}$$

Set $\underline{x}_q=X_{(m)}+z_{\alpha/2}\,\hat{a}$, $\bar{x}_q=X_{(m)}+z_{1-(\alpha/2)}\,\hat{a}$ and $I=[\underline{x}_q,\bar{x}_q]$. By Eq. (2.3)

$$P_{\tau,a}(x_q\notin I)=\alpha \text{ for } -\infty<\tau<\infty \text{ and } a>0 \quad .$$

In other words, $\bar{I}$ is a classical $100(1-\alpha)\%$ confidence interval for x_q within the context of the two-parameter exponential model. It is called the <u>two-parameter exponential</u> confidence interval procedure. The procedure is location and scale invariant. That is, if $X_{(1)},\ldots,X_{(m)}$ are replaced by $d+bX_{(1)},\ldots,d+bX_{(m)}$ with $b>0$, then the left and right end-points $\underline{x}_q$ and $\bar{x}_q$ of $\bar{I}$ are replaced by $d+b\underline{x}_q$ and $d+b\bar{x}_q$, respectively.

Formula (2.3) and hence the confidence interval procedure $\bar{I}$ have a generalized Bayesian interpretation. To see this, let π denote the generalized prior density on $(-\infty,\infty)x(0,\infty)$ defined by $\pi(\tau,a)=1/a$. The corresponding posterior density is defined by

$$\pi(\tau,a|x_1,\ldots,x_m)$$

$$= \frac{\pi(\tau,a)f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;\tau,a)}{\iint\pi(\tau,a)f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;\tau,a)\,d\tau\,da}$$

$$= \frac{a^{1-m}e^{-\sum_1^m(x_i-x_m)/a}\,e^{-m(x_m-\tau)/a}\left[1-e^{-(x_m-\tau)/a}\right]^{n-m}}{\int_0^\infty a^{1-m}e^{-\sum_1^m(x_i-x_m)/a}\,da\int_{-\infty}^{x_m}e^{-m(x_m-\tau)/a}\left[1-e^{-(x_m-\tau)/a}\right]^{n-m}\,d\tau}$$

for $\tau \leq x_m$ and

$$\pi(\tau,a|x_1,\ldots,x_m) = 0 \qquad \text{for } \tau > x_m \quad .$$

(It is understood here that $x_1>\ldots>x_m$.)

Let $\underset{\sim}{\tau}$ and $\underset{\sim}{a}$ denote random variables having joint density $\pi(\tau,a|x_1,\ldots,x_m)$. Then the marginal density $\pi_{\underset{\sim}{a}}(a|x_1,\ldots,x_m)$ of $\underset{\sim}{a}$ is given by

$$\pi_{\underset{\sim}{a}}(a|x_1,\ldots,x_m)$$

$$= \frac{\int_{-\infty}^{\infty}\pi(\tau,a|x_1,\ldots,x_m)\,d\tau}{\int_0^{\infty}da\int_{-\infty}^{\infty}\pi(\tau,a|x_1,\ldots,x_m)\,d\tau}$$

$$= \frac{\hat{a}}{a^2}\,g\left(\frac{\hat{a}}{a};m-1,m-1\right) \quad ,$$

where $\hat{a} = \sum_1^m(x_i-x_m)/(m-1)$ and $g(\cdot;m-1,m-1)$ is a gamma density as defined above. The conditional density $\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|a;x_1,\ldots,x_m)$ of $\underset{\sim}{\tau}$ given $\underset{\sim}{a} = a$ is obtained as

$$\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|a;x_1,\ldots,x_m)$$

$$= \frac{\pi(\tau,a|x_1,\ldots,x_m)}{\int_{-\infty}^{\infty}\pi(\tau,a|x_1,\ldots,x_m)\,d\tau}$$

$$= \frac{n!}{(m-1)!(n-m)!}\,a^{-1}\,e^{-m(x_m-\tau)/a}\left[1-e^{-(x_m-\tau)/a}\right]^{n-m}$$

for $\tau < x_m$ and $\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|\mu;x_1,\ldots,x_m) = 0$ for $\tau > x_m$.

Set $\underset{\sim}{p} = e^{-(x_m-\underset{\sim}{\tau})/\underset{\sim}{a}}$ and note that $\underset{\sim}{\tau} = x_m + \underset{\sim}{a} \log \underset{\sim}{p}$. The conditional density $\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;x_1,\ldots,x_m)$ of $\underset{\sim}{p}$ given $\underset{\sim}{a} = a$ is obtained as

$$\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|\mu;x_1,\ldots,x_m)$$

$$= \frac{a}{p}\, \pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(x_m+a \log p|a;x_1,\ldots,x_m)$$

$$= \frac{n!}{(m-1)!(n-m)!}\, p^{m-1}(1-p)^{n-m} \qquad \text{for } 0<p<1$$

and $\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;x_1,\ldots,x_m)=0$ otherwise. Consequently the conditional distribution function of $\underset{\sim}{p}$ given $\underset{\sim}{a} = a$ is obtained as

$$F_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;x_1,\ldots,x_m) = B(p;m,n-m+1) \quad .$$

For $0<q<1$ set

$$\underset{\sim}{x}_q = \underset{\sim}{\tau} + \underset{\sim}{a} \log(1/q) = x_m + \underset{\sim}{a} \log(\underset{\sim}{p}/q) \quad .$$

Then the conditional distribution function of x_q given $\underset{\sim}{a} = a$ is obtained as

$$F_{\underset{\sim}{x}_q|\underset{\sim}{a}}(x|a;x_1,\ldots,x_m)$$

$$= F_{\underset{\sim}{p}|\underset{\sim}{a}}\left(q\, e^{(x-x_m)/a}|a;x_1,\ldots,x_m\right)$$

$$= B\left(q\, e^{(x-x_m)/a};m,n-m+1\right) \quad .$$

Consequently the distribution function of $\underset{\sim}{X}_q$ is given by

$$F_{\underset{\sim}{X}_q}(x|x_1,\ldots,x_m)$$

$$= \int_0^\infty F_{\underset{\sim}{X}_q|\underset{\sim}{a}}(x|a;x_1,\ldots,x_m)\pi_{\underset{\sim}{a}}(a|x_1,\ldots,x_m)\,da$$

$$= \int_0^\infty B\left(q\, e^{(x-x_m)/a};m,n-m+1\right)\frac{\hat{a}}{a^2}\, g\left(\frac{\hat{a}}{a};m-1,m-1\right)\,da$$

$$= \int_0^\infty B\left(q\, e^{(x-x_m)t/\hat{a}};m,n-m+1\right)g(t;m-1,m-1)\,dt \quad .$$

Therefore

$$F_{\underset{\sim}{X}_q}(x_m+z_\alpha\hat{a}|x_1,\ldots,x_m) = \alpha \quad ,$$

where z_α is defined by Eq. (2.2). This yields the generalized Bayesian interpretation of Eq. (2.3).

2.1.1 Modifications to Handle Grouped Data

In many applications $X_1,\ldots,X_n$ are rounded up or down or grouped to yield a small to moderate number of distinct values. This rounding or grouping can have an adverse effect on the confidence interval procedure described above unless the procedure is appropriately modified.

Specifically let $k\geq 2$ denote a positive integer and let $-\infty=d_0<d_1<\ldots<d_{k-1}<d_k=\infty$. For $1\leq j<k$ let N_j denote the number of sample values in the interval (d_{j-1},d_j). Then

$$P_{\tau,a}(N_1=n_1,\ldots,N_k=n_k)$$

$$= C(n_1,\ldots,n_k)\prod_{j=1}^{k}[F(d_j;\tau,a) - F(d_{j-1};\tau,a)]^{n_j} \quad ,$$

where $F(-\infty;\tau,a)=0$, $F(\infty;\tau,a)=1$, $n_1,\ldots,n_k$ are nonnegative integers adding up to n and

$$C(n_1,\ldots,n_k) = \frac{n!}{n_1!\cdot\ldots\cdot n_k!} \quad .$$

If $n_1>0$, then

$$P_{\tau,a}(N_1=n_1,\ldots,N_k=n_k)$$

$$= C(n_1,\ldots,n_k)\left[1-e^{-(d_1-\tau)/a}\right]^{n_1}\prod_{j=2}^{k}\left[e^{-(d_{j-1}-\tau)/a} - e^{-(d_j-\tau)/a}\right]^{n_j}$$

$$= C(n_1,\ldots,n_k)\prod_{j=2}^{k}\left[e^{-(d_{j-1}-d_1)/a} - e^{-(d_j-d_1)/a}\right]^{n_j} e^{-(n-n_1)(d_1-\tau)/a}\left[1-e^{-(d_1-\tau)/a}\right]^{n_1}$$

for $\tau < d_1$ and $a > 0$, while the indicated probability equals zero for $\tau \geq d_1$.

An approximate 100 $(1-\alpha)$% confidence interval for x_q will now be obtained by modifying the generalized Bayes derivation of $\bar{I}$ given above. Let π again denote the generalized prior density on $(-\infty,\infty)x(0,\infty)$ defined by $\pi(\tau,a)=1/a$. Suppose $1\leq n_1<n$. Consider the corresponding posterior density defined by

$$\pi(\tau,a|n_1,\ldots,n_k)$$

$$= \frac{\pi(\tau,a)P_{\tau,a}(N_1=n_1,\ldots,N_k=n_k)}{\iint[\pi(\tau,a)P_{\tau,a}(N_1=n_1,\ldots,N_k=n_k)]\,d\tau\,da}$$

$$= \frac{a^{-1}\prod_{j=2}^{k}\left[e^{-(d_{j-1}-d_1)/a}-e^{-(d_j-d_1)/a}\right]^{n_j} e^{-(n-n_1)(d_1-\tau)/a}\left[1-e^{-(d_1-\tau)/a}\right]^{n_1}}{\iint a^{-1}\prod_{j=2}^{k}\left[e^{-(d_{j-1}-d_1)/a}-e^{-(d_j-d_1)/a}\right]^{n_j} e^{-(n-n_1)(d_1-\tau)/a}\left[1-e^{-(d_1-\tau)/a}\right]^{n_1} d\tau\,da}$$

for $\tau \le d_1$ and $a > 0$, while $\pi(\tau,a|n_1,\ldots,n_k) = 0$ otherwise. Observe that

$$\int_{-\infty}^{d_1} e^{(n-n_1)(d_1-\tau)/a}\left[1-e^{-(d_1-\tau)/a}\right]^{n_1} d\tau$$

$$= a\int_0^{\infty} e^{-(n-n_1)\tau}(1-e^{-\tau})^{n_1}\,d\tau$$

$$= a\int_0^1 p^{n-n_1-1}(1-p)^{n_1}\,dp = a\,\frac{n_1!(n-n_1-1)!}{n!}\quad .$$

Let $\underset{\sim}{\tau}$ and $\underset{\sim}{a}$ denote random variables having joint density $\pi(\tau,a|n_1,\ldots,n_k)$. Then the marginal density $\pi_{\underset{\sim}{a}}(a|n_1,\ldots,n_k)$ is given by

$$\pi_{\underset{\sim}{a}}(a|n_1,\ldots,n_k) = \frac{\prod_{j=2}^{k}\left[e^{-(d_{j-1}-d_1)/a}-e^{-(d_j-d_1)/a}\right]^{n_j}}{\int_0^{\infty}\prod_{j=2}^{k}\left[e^{-(d_{j-1}-d_1)/a}-e^{-(d_j-d_1)/a}\right]^{n_j} da}\quad . \tag{2.4}$$

The conditional density $\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|a;n_1,\ldots,n_k)$ of $\underset{\sim}{\tau}$ given $\underset{\sim}{a} = a$ is obtained as

$$\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|a;n_1,\ldots,n_k)$$

$$= \frac{n!}{a\, n_1!(n-n_1-1)!}\; e^{-(n-n_1)(d_1-\tau)/a}\left[1 - e^{-(d_1-\tau)/a}\right]^{n_1}$$

for $\tau \leq d_1$, and $\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(\tau|a;n_1,\ldots,n_k) = 0$ for $\tau > d_1$.

Set $\underset{\sim}{p} = e^{-(d_1-\underset{\sim}{\tau})/\underset{\sim}{a}}$ and note that $\underset{\sim}{\tau} = d_1 + \underset{\sim}{a} \log \underset{\sim}{p}$. The conditional density $\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;n_1,\ldots,n_k)$ of $\underset{\sim}{p}$ given $\underset{\sim}{a} = a$ is obtained as

$$\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;n_1,\ldots,n_k)$$

$$= \frac{a}{p}\,\pi_{\underset{\sim}{\tau}|\underset{\sim}{a}}(d_1 + a \log p|a;n_1,\ldots,n_k)$$

$$= \frac{n!}{n_1!(n-n_1-1)!}\, p^{n-n_1-1}(1-p)^{n_1} \quad \text{for } 0 < p < 1$$

and $\pi_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;n_1,\ldots,n_k) = 0$ otherwise.

The conditional distribution function of $\underset{\sim}{p}$ given $\underset{\sim}{a} = a$ is obtained as

$$F_{\underset{\sim}{p}|\underset{\sim}{a}}(p|a;n_1,\ldots,n_k) = B(p;n-n_1,n_1+1)$$

where $B(p;\alpha,\beta)$ is as defined above.

For $0 < q < 1$ set

$$\underset{\sim}{x}_q = \underset{\sim}{\tau} + \underset{\sim}{a} \log (1/q) = d_1 + \underset{\sim}{a} \log (\underset{\sim}{p}/q) \quad .$$

Then the conditional distribution function of $\underset{\sim}{x}_q$ given $\underset{\sim}{a} = a$ is obtained as

$$\begin{aligned} & F_{\underset{\sim}{x}_q|\underset{\sim}{a}}(x|a;n_1,\ldots,n_k) \\ &= F_{\underset{\sim}{p}|\underset{\sim}{a}}\left(q\, e^{(x-d_1)/a} \,|a;n_1,\ldots,n_k\right) \\ &= B\left(q\, e^{(x-d_1)/a};n-n_1,n_1+1\right) \quad . \end{aligned}$$

Consequently the distribution function of $\underset{\sim}{x}_q$ is given by

$$\begin{aligned} & F_{\underset{\sim}{x}_q}(x|n_1,\ldots,n_k) \\ &= \int_0^\infty F_{\underset{\sim}{x}_q|\underset{\sim}{a}}(x|a;n_1,\ldots,n_k)\ \pi_{\underset{\sim}{a}}(a|n_1,\ldots,n_k)\ da \\ &= \int_0^\infty B\left(q\, e^{(x-d_1)/a};n-n_1,n_1+1\right) \pi_{\underset{\sim}{a}}(a|n_1,\ldots,n_k)\ da \quad , \end{aligned}$$

where $\pi_{\underset{\sim}{a}}(a|n_1,\ldots,n_k)$ is given explicitly in Eq. (2.4).

Given $0 < \alpha < 1$, define $\underline{x}_q = \underline{x}_q(n_1,\ldots,n_k)$ and $\overline{x}_q = \overline{x}_q(n_1,\ldots,n_k)$ by

$$F_{\underset{\sim}{x}_q}(\underline{x}_q|n_1,\ldots,n_k) = \frac{\alpha}{2} \text{ and } F_{\underset{\sim}{x}_q}(\overline{x}_q|n_1,\ldots,n_k) = 1 - \frac{\alpha}{2} \quad .$$

Also set

$$\bar{I} = [\underline{x}_q(N_1,\ldots,N_k),\ \bar{x}_q(N_1,\ldots,N_k)] \quad .$$

Then $\bar{I}$ should be a good substitute for the two-parameter exponential 100 $(1-\alpha)$% confidence interval when the sample data are rounded or grouped. (Note that additional rounding or grouping may be required to guarantee that $0 < N_1 < n$. This provides no problem in practice, for if the requirement cannot be met, the original rounded or grouped observations are identical and there are no reasonable confidence intervals for x_q.)

2.2 TWO-PARAMETER WEIBULL PROCEDURE

Consider the two-parameter Weibull model

$$F(x;t,\beta) = 1 - e^{-tx^{\beta}} \quad , \qquad x > 0 \quad ,$$

$$= 0 \quad , \qquad x \leq 0 \quad ,$$

where $t > 0$ and $\beta > 0$. Correspondingly

$$S(x;t,\beta) = e^{-tx^{\beta}} \quad , \qquad x > 0 \quad ,$$

$$= 0 \quad , \qquad x \leq 0 \quad ,$$

and

$$x_q = t^{-1/\beta}\left(\log \frac{1}{q}\right)^{1/\beta} \quad , \qquad 0 < q \leq 1 \quad .$$

In this model β is a shape parameter and $a = t^{-1/\beta}$ is a scale parameter. Alternatively log a and $1/\beta$ are respectively the location and scale parameters for the distribution of log X, where X has distribution function $F(\cdot;t,\beta)$.

Given $2 \leq m \leq n$, let $X_{(1)},\ldots,X_{(m)}$ denote the upper m order statistics based on a random sample of size n from $F(\cdot;t,\beta)$. The joint density of these random variables is given by

$$f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;t,\beta)$$

$$= \frac{n!}{(n-m)!}\beta^m\left(\prod_1^m x_i\right)^{\beta-1} t^m\left(1 - e^{-tx_m^\beta}\right)^{n-m} e^{-t\sum_1^m x_i^\beta} .$$

The maximum likelihood estimators of t and β are not easily found.

An approximate generalized Bayes confidence interval procedure for x_q will now be obtained. Let π denote the generalized prior density on $(0,\infty)x(0,\infty)$ defined by $\pi(t,\beta) = 1/t\beta$. The corresponding posterior density is defined by

$$\pi(t,\beta|x_1,\ldots,x_m)$$

$$= \frac{\pi(t,\beta)f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;t,\beta)}{\iint\pi(t,\beta)f_{X_{(1)},\ldots,X_{(m)}}(x_1,\ldots,x_m;t,\beta)\,dt\,d\beta}$$

$$= \frac{\beta^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} t^{m-1}\left(1 - e^{-tx_m^\beta}\right)^{n-m} e^{-t\sum_1^m x_i^\beta}}{\int_0^\infty \beta^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} d\beta \int_0^\infty t^{m-1}\left(1 - e^{tx_m^\beta}\right)^{n-m} e^{-t\sum_1^m x_i^\beta}\,dt} .$$

For $\nu > 0$ and $t > 0$ set

$$g(\nu,t) = t^{m-1}(1 - e^{-\nu t})^{n-m} e^{-mt} \quad .$$

Also set

$$\hat{t} = \hat{t}(\beta;x_1,\ldots,x_m) = \frac{1}{\frac{1}{m}\sum_1^m x_i^\beta} \quad .$$

Then

$$\pi(t,\beta|x_1,\ldots,x_m)$$

$$= \frac{(\hat{t}\beta)^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} g(\hat{t}x_m^\beta, t/\hat{t})}{\int_0^\infty (\hat{t}\beta)^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} d\beta \int_0^\infty g(\hat{t}x_m^\beta, t/\hat{t})\, dt} \quad .$$

Let $\underset{\sim}{t}$ and $\underset{\sim}{\beta}$ denote random variables having joint density $\pi(t,\beta|x_1,\ldots,x_m)$. Then the marginal density $\pi_{\underset{\sim}{\beta}}(\beta|x_1,\ldots,x_m)$ of $\underset{\sim}{\beta}$ is given by

$$\pi_{\underset{\sim}{\beta}}(\beta|x_1,\ldots,x_m)$$

$$= \frac{\int_0^\infty \pi(t,\beta|x_1,\ldots,x_m)\ dt}{\int_0^\infty d\beta \int_0^\infty \pi(t,\beta|x_1,\ldots,x_m)\ dt}$$

$$= \frac{(\hat{t}\beta)^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} \int_0^\infty g(\hat{t}x_m^\beta, t/\hat{t})\ dt}{\int_0^\infty (\hat{t}\beta)^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} d\beta \int_0^\infty g(\hat{t}x_m^\beta, t/\hat{t})\ dt} \quad .$$

The conditional density $\pi_{\underset{\sim}{t}|\underset{\sim}{\beta}}(t|\beta;x_1,\ldots,x_m)$ of $\underset{\sim}{t}$ given $\underset{\sim}{\beta} = \beta$ is obtained as

$$\pi_{\underset{\sim}{t}|\underset{\sim}{\beta}}(t|\beta;x_1,\ldots,x_m)$$

$$= \frac{\pi(t,\beta|x_1,\ldots,x_m)}{\int_0^\infty \pi(t,\beta|x_1,\ldots,x_m)\ dt}$$

$$= \frac{g(\hat{t}x_m^\beta, t/\hat{t})}{\int_0^\infty g(\hat{t}x_m^\beta, t/\hat{t})\ dt} \quad .$$

The corresponding conditional distribution function of $\underset{\sim}{t}$ given $\underset{\sim}{\beta} = \beta$ is obtained as

$$F_{\underset{\sim}{t}|\underset{\sim}{\beta}}(t|\beta;x_1,\ldots,x_m)$$

$$= \int_0^t \pi_{\underset{\sim}{t}|\underset{\sim}{\beta}}(u|\beta;x_1,\ldots,x_m)\, du$$

$$= \frac{\int_0^t g(\hat{t}\, x_m^{\beta},\ u/\hat{t})\, du}{\int_0^{\infty} g(\hat{t}\, x_m^{\beta},\ u/\hat{t})\, du} \quad .$$

For $0 < q < 1$ set

$$\underset{\sim}{x}_q = \underset{\sim}{t}^{-1/\underset{\sim}{\beta}} \left(\log \frac{1}{q}\right)^{1/\underset{\sim}{\beta}} \quad .$$

Then $\underset{\sim}{x}_q \leq x$ if and only if

$$\underset{\sim}{t} \geq x^{-\underset{\sim}{\beta}} \log (1/q) \quad .$$

Thus the conditional distribution function of $\underset{\sim}{x}_q$ given $\underset{\sim}{\beta} = \beta$ is obtained as

$$F_{\underset{\sim}{x}_q|\underset{\sim}{\beta}}(x|\beta;x_1,\ldots,x_m)$$

$$= 1 - F_{\underset{\sim}{t}|\underset{\sim}{\beta}}\left(x^{-\beta} \log (1/q)|\beta;x_1,\ldots,x_m\right) \quad .$$

The distribution function of $\underset{\sim}{x}_q$ can be expressed in terms of quantities defined above according to the formula

$$F_{\underset{\sim}{x}_q}(x|x_1,\ldots,x_m)$$

$$= \int_0^\infty F_{\underset{\sim}{x}_q|\underset{\sim}{\beta}}(x|\beta;x_1,\ldots,x_m)\ \pi_{\underset{\sim}{\beta}}(\beta|x_1,\ldots,x_m)\ d\beta \quad .$$

Given $0 < \alpha < 1$ defined $\underline{x}_q$ and $\overline{x}_q$ in terms of $X_{(1)},\ldots,X_{(m)}$ according to the formulas $F_{\underset{\sim}{x}_q}(\underline{x}_q|X_{(1)},\ldots,X_{(m)}) = \alpha/2$ and $F_{\underset{\sim}{x}_q}(\overline{x}_q|X_{(1)},\ldots,X_{(m)}) = 1 - \alpha/2$. Also set $\overline{I} = [\underline{x}_q,\overline{x}_q]$. Then $P_{t,\beta}(x_q \notin \overline{I}) = \alpha$ for all $t > 0$ and $\beta > 0$. The proof of this result, which will not be given here, depends on the observation made at the beginning of this subsection that the two-parameter Weibull model can be viewed as a location-scale model. The result shows that $\overline{I}$ is a classical $100(1-\alpha)\%$ confidence interval for x_q within the context of the model. It is called the <u>two-parameter Weibull</u> confidence interval procedure. The procedure is invariant under scale and power transformations. That is, if $X_{(1)},\ldots,X_{(m)}$ are replaced by $d\ X_{(1)}^b,\ldots,d\ X_{(m)}^b$ respectively with $d > 0$ and $b > 0$, the left and right end-points $\underline{x}_q$ and $\overline{x}_q$ of $\overline{I}$ are replaced by $d\ \underline{x}_q^b$ and $d\ \overline{x}_q^b$ respectively. [To see this, observe first that

$$\pi(t,\beta|d\ x_1^b,\ldots,d\ x_m^b) = d^\beta\ b\pi(d^\beta t, b\beta|x_1,\ldots,x_m)$$

and then use this equation to show that

$$F_{\underset{\sim}{x}_q}(d\ x^b|d\ x_1^b,\ldots,d\ x_m^b) = F_{\underset{\sim}{x}_q}(x|x_1,\ldots,x_m) \quad .]$$

The confidence intervals for x_q can be transformed as described in Subsection 2.1 to yield $100(1-\alpha)\%$ confidence intervals for $S(x)$.

Unfortunately the two-parameter Weibull confidence interval procedure is not computationally feasible. To obtain an approximate version of the procedure which is computationally feasible, for $\nu > 0$ and $t > 0$ set $h(\nu,t) = \log g(\nu,t) = (m-1) \log t + (n-m) \log (1-e^{-\nu t}) - mt$ and let $h'(\nu,t)$, etc., denote differentiation of h with respect to t. Then

$$h'(\nu,t) = \frac{m-1}{t} + \frac{(n-m)\nu}{e^{\nu t}-1} - m \quad ,$$

$$h''(\nu,t) = -\frac{m-1}{t^2} - \frac{(n-m)\nu^2 e^{\nu t}}{(e^{\nu t}-1)^2} < 0 \quad ,$$

and

$$h'''(\nu,t) = \frac{2(m-1)}{t^3} + \frac{(n-m)\nu^3 e^{\nu t}(e^{\nu t}+1)}{(e^{\nu t}-1)^3} > 0 \quad .$$

Thus $h(\nu,t)$ has a unique maximum at $t_0 = t(\nu)$ which is the unique root of $h'(\nu,t) = 0$. This root can be found by applying Newton's method to the function $h'(\nu,\cdot)$.

Consider the approximation

$$h(\nu,t) \doteq h(\nu,t_0) + \frac{1}{2} h''(\nu,t_0)(t-t_0)^2$$

$$= \log g(\nu,t_0) + \frac{1}{2} h''(\nu,t_0)(t-t_0)^2 \quad .$$

Correspondingly

$$g(\nu,t) \doteq g(\nu,t_0)\, e^{h''(\nu,t_0)(t-t_0)^2/2} \quad .$$

This can be written as

$$g(\nu,t) \doteq C(\nu)N(t;t(\nu),\sigma^2(\nu)) \quad ,$$

where

$$\sigma(\nu) = \sqrt{-1/h''(\nu,t(\nu))} \quad ,$$

$$C(\nu) = g(\nu,t(\nu))\sigma(\nu)\sqrt{2\pi} \quad ,$$

and $N(\cdot;\mu,\sigma^2)$ is the normal density with mean μ and variance σ^2. The approximation for g in turn yields the approximations

$$\pi_{\beta}(\beta|x_1,\ldots,x_m) \doteq \frac{\hat{t}^m\, \beta^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} C(\hat{\nu}(\beta))}{\int_0^\infty \hat{t}^m\, \beta^{m-1}\left(\prod_1^m x_i\right)^{\beta-1} C(\hat{\nu}(\beta))\, d\beta}$$

and

$$\pi_{t|\beta}(t|\beta;x_1,\ldots,x_m) \doteq N\left(t;\hat{t}\ t(\hat{\nu}(\beta)),\ \hat{t}^2\ \sigma^2(\hat{\nu}(\beta))\right) \quad ,$$

where $\hat{\nu}(\beta) = \hat{t}\ x_m^{\beta}$.

Let Φ denote the standard normal distribution function and set $Q = 1 - \Phi$. Then

$$F_{\underset{\sim}{t}|\underset{\sim}{\beta}}(t|\beta;x_1,\ldots,x_m) \doteq \Phi\left(\frac{\hat{t}^{-1}t - t(\hat{\nu}(\beta))}{\sigma(\hat{\nu}(\beta))}\right) \quad ,$$

so

$$F_{\underset{\sim}{x_q}|\underset{\sim}{\beta}}(x|\beta;x_1,\ldots,x_m) \doteq Q\left(\frac{\hat{t}^{-1}x^{-\beta}\log(1/q) - t(\hat{\nu}(\beta))}{\sigma(\hat{\nu}(\beta))}\right) \quad .$$

Set

$$\dot{F}_{\underset{\sim}{x_q}}(x|x_1,\ldots,x_m)$$

$$= \int_0^\infty \dot{F}_{\underset{\sim}{x_q}|\underset{\sim}{\beta}}(x|\beta;x_1,\ldots,x_m)\ \dot{\pi}_{\underset{\sim}{\beta}}(\beta|x_1,\ldots,x_m)\ d\beta \quad ,$$

where $\dot{\pi}_\beta$ and $\dot{F}_{\underset{\sim}{x_q}|\underset{\sim}{\beta}}$ are the approximations to π_β and $F_{\underset{\sim}{x_q}|\underset{\sim}{\beta}}$ just determined.

Given $0 < \alpha < 1$ define $\underline{x}_q$ and $\overline{x}_q$ in terms of $X_{(1)},\ldots,X_{(m)}$ according to the formulas

$$\dot{F}_{\underset{\sim}{x_q}}\left(\underline{x}_q|X_{(1)},\ldots,X_{(m)}\right) = \frac{\alpha}{2} \text{ and } F_{\underset{\sim}{x_q}}\left(\overline{x}_q|X_{(1)},\ldots,X_{(m)}\right) = 1 - \frac{\alpha}{2} \quad .$$

Then $\overline{I} = [\underline{x}_q, \overline{x}_q]$ determines the approximate two-parameter Weibull confidence interval procedure. It is also invariant under scale and power transformations.

2.3 QUADRATIC TAIL PROCEDURE

Let $3 \leq m \leq n$ and let $X_{(1)},\ldots,X_{(m)}$ denote the upper m order statistics based on a random sample of size n from F. Set $\ell(x) = -\log S(x) = -\log [1-F(x)]$. Then $S(x) = e^{-\ell(x)}$, so that

$$S(x) = S(X_{(m)})\, e^{-[\ell(x)-\ell(X_{(m)})]}, \quad x > X_{(m)} \quad .$$

In particular, if $x_q \geq X_{(m)}$, then

$$q = S(X_{(m)})\, e^{-[\ell(x_q)-\ell(X_{(m)})]}$$

and hence

$$\ell(x_q) - \ell(X_{(m)}) = \log \frac{S(X_{(m)})}{q} \quad .$$

In other words, x_q is the solution to the equation

$$\ell(x) - \ell(X_{(m)}) = y \quad , \tag{2.5}$$

where $y = \log [S(X_{(m)})/q]$. The solution to Eq. (2.5) can be written as

$$x = X_{(m)} + L(y) \quad , \tag{2.6}$$

where L depends on $X_{(m)}$ as well as the (unknown) distribution function F. By Eq. (2.6)

$$x_q = X_{(m)} + L\left(\log \frac{S(X_{(m)})}{q}\right) \quad , \quad 0 < q \leq S(X_{(m)}) \quad . \tag{2.7}$$

This suggests estimates $\hat{x}_q$ of x_q having the form

$$\hat{x}_q = X_{(m)} + \hat{L}\left(\log \frac{\hat{S}(X_{(m)})}{q}\right), \quad 0 < q \leq \hat{S}(X_{(m)}) \quad .$$

It is natural to estimate $S(X_{(m)})$ by $\hat{S}(X_{(m)}) = m/n$. This leads to

$$\hat{x}_q = X_{(m)} + \hat{L}\left(\log \frac{m}{nq}\right), \quad 0 < q \leq \frac{m}{n} \quad . \tag{2.8}$$

Consider for example a distribution function F belonging to the two-parameter exponential model. Then

$$F(x) = 1 - e^{-(x-\tau)/a}, \quad x \geq \tau \quad ,$$

where $-\infty < \tau < \infty$ and $a > 0$. Correspondingly,

$$\ell(x) = \frac{x-\tau}{a}, \quad x \geq \tau \quad ,$$

and

$$L(y) = ay, \quad y \geq 0 \quad .$$

If a is estimated by

$$\hat{a} = \frac{1}{m-1} \sum_{1}^{m-1} [X_{(i)} - X_{(m)}]$$

and $L(y) = ay$ by

$$\hat{L}(y) = \hat{a}y \quad , \quad y \geq 0 \quad , \tag{2.9}$$

Eq. (2.8) reduces to the exponential tail estimator described in Subsection 2.1.

A natural extension of Eq. (2.9) is estimators of L having the form

$$\hat{L}(y) = \hat{a}y + \frac{\hat{b}}{2} y^2 \quad , \quad y \geq 0 \quad , \tag{2.10}$$

where the additional term $\hat{b}\ y^2/2$ hopefully properly takes into account small to moderate departures of the tail of F from the two-parameter exponential model. Together, Eqs. (2.8) and (2.10) yield the estimator.

$$\hat{x}_q + X_{(m)} + \hat{a} \log \frac{m}{nq} + \frac{\hat{b}}{2}\left(\log \frac{m}{nq}\right)^2 \quad , \quad 0 < q \leq \frac{m}{n} \quad . \tag{2.11}$$

It is desirable that $\hat{x}_q$ be a nondecreasing function of $1/q$. This is true for the estimator given in Eq. (2.11) if $\hat{a}$ and $\hat{b}$ are nonnegative. Otherwise the estimator can be modified in an obvious way to make it nondecreasing in $1/q$ [set $\hat{x}_q = \hat{x}_{q_o}$ for $q \leq q_o$, where q_o is chosen as large as possible subject to the constraint that $\hat{L}$ given by Eq. (2.10) is nondecreasing in y for $0 \leq y \leq \log (m/nq_o)$].

In order to determine specific choices of the quantities $\hat{a}$ and $\hat{b}$ appearing in the definition of L, it will be assumed that

$$L(y) = ay + \frac{b}{2} y^2 \quad , \quad y \geq 0 \quad , \tag{2.12}$$

where $a \geq 0$ and $-\infty < b < \infty$. Of course this can be exactly true only if $b \geq 0$, so the following discussion is "formal" if $b < 0$.

By Eqs. (2.7) and (2.12)

$$x_q = X_{(m)} + a \log \frac{S(X_{(m)})}{q} + \frac{b}{2}\left[\log \frac{S(X_{(m)})}{q}\right]^2 \quad .$$

If m is reasonably large, then $m/n = \hat{S}(X_{(m)}) \doteq S(X_{(m)})$ and hence

$$x_q \doteq X_{(m)} + a \log \frac{m}{nq} + \frac{b}{2}\left(\log \frac{m}{nq}\right)^2 \quad . \tag{2.13}$$

Set

$$L = \log \frac{m}{nq} \text{ and } N = \frac{1}{2} \log \frac{m}{nq} \quad .$$

Also let

$$\hat{\theta} = \hat{a} + N\hat{b}$$

be considered as an estimator of

$$\theta = a + Nb \quad .$$

Then Eqs. (2.13) and (2.11) can be rewritten, respectively, as

$$x_q \doteq X_{(m)} + L\theta \tag{2.14}$$

and

$$\hat{x}_q = X_{(m)} + L\hat{\theta} \quad . \tag{2.15}$$

Similarly a confidence interval $\overline{J} = [\underline{\theta}, \overline{\theta}]$ for θ yields a confidence interval $\overline{I} = [\underline{x}_q, \overline{x}_q]$ for x_q, where $\underline{x}_q = X_{(m)} + L\underline{\theta}$ and $\overline{x}_q = X_{(m)} + L\overline{\theta}$.

It is natural to consider estimators $\hat{\theta}$ of θ which are linear combinations of $X_{(k)} - X_{(m)}$, $1 \leq k \leq m$, with constant coefficients. Such an estimator can be written in the form

$$\hat{\theta} = \sum_{1}^{m-1} k\, \omega_k [X_{(k)} - X_{(k+1)}] \quad . \tag{2.16}$$

It will be shown in Appendix I that the expected value of such an estimator $\hat{\theta}$ is given by

$$E\,\hat{\theta} = \left(\sum_{1}^{m-1} \omega_k\right) a + \left(\sum_{1}^{m-1} \mu_k\, \omega_k\right) b \quad , \tag{2.17}$$

where

$$\mu_k = \sum_{j=k}^{m-1} \frac{1}{j} \quad .$$

Given an integer J such that $2 \leq J \leq m$, set

$$\bar{S}_J = \frac{1}{J-1} \sum_{1}^{J-1} [X_{(k)} - X_{(J)}] = \frac{1}{J-1} \sum_{1}^{J-1} k[X_{(k)} - X_{(k+1)}] \quad .$$

By Eq. (2.17)

$$E\,\bar{S}_J = a + \left(\frac{1}{J-1} \sum_{1}^{J-1} \mu_k\right) b \quad .$$

It is easily verified that

$$\frac{1}{J-1}\sum_{1}^{J-1}\mu_k = 1 + \mu_J \quad .$$

Consequently

$$E\,\bar{S}_J = a + (1 + \mu_J)b \quad . \tag{2.18}$$

It is well known that

$$\lim_{m\to\infty}\sum_{j=1}^{m-1}\frac{1}{j} - \log(m-1) = \gamma \quad ,$$

where γ is Euler's constant. Thus for large values of m and $J \leq m$

$$\mu_J \doteq \log\frac{m-1}{J-1} \quad ;$$

so $\bar{S}_J$ is an approximately unbiased estimator of θ if

$$1 + \log\frac{m-1}{J-1} \doteq N$$

or, equivalently, if

$$J - 1 \doteq (m-1)e^{1-N} \quad . \tag{2.19}$$

Suppose from now on that Eq. (2.19) holds. It is shown in Appendix I that

$$\mathrm{Var}(\bar{S}_J) = \frac{1}{J-1}\,[a + (1+N)b]^2 + (1+\lambda_J)b^2 \qquad (2.20)$$

where

$$\lambda_J \doteq 1 - \frac{J-1}{m-1} \quad .$$

Equivalently

$$\mathrm{Var}(\bar{S}_J) = \frac{1}{J-1}\,[(\theta+b)^2 + (1+\lambda_J)b^2] \quad . \qquad (2.21)$$

Suppose now that $(\bar{S}_J - \theta)/\mathrm{SD}(\bar{S}_J)$ has approximately the standard normal distribution. Given $0 < \alpha < 1$ choose $z_{\alpha/2}$ such that

$$\int_{z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx = \frac{\alpha}{2} \quad .$$

Then

$$P\left(-z_{\alpha/2}\,\mathrm{SD}(\bar{S}_J) \le \bar{S}_J - \theta \le z_{\alpha/2}\,\mathrm{SD}(\bar{S}_J)\right) \doteq 1 - \alpha \quad . \qquad (2.22)$$

The inequality inside the probability in Eq. (2.22) can be written as

$$(\bar{S}_J - \theta)^2 \le z_{\alpha/2}^2\,\mathrm{Var}(\bar{S}_J) \quad . \qquad (2.23)$$

Set $\gamma = z_{\alpha/2}/\sqrt{J-1}$. By Eq. (2.21), Eq. (2.23) can be rewritten as

$$(\bar{S}_J - \theta)^2 \leq \gamma^2[(\theta+b) + (1+\lambda_J)b^2] \quad .$$

This inequality can be solved to yield the interval $\theta^- \leq \theta \leq \theta^+$ where

$$\theta^{\pm} = \frac{\bar{S}_J + \gamma^2 b \pm \gamma\sqrt{(\bar{S}_J+b)^2 + b^2(1-\gamma^2)(1+\lambda_J)}}{1 - \gamma^2} \quad . \tag{2.24}$$

In order to get calculable estimates it is necessary to replace b in Eq. (2.24) by an estimator $\hat{b}$. By Eq. (2.18)

$$\hat{b} = \frac{\bar{S}_J - \bar{S}_m}{\mu_J} \tag{2.25}$$

is an unbiased estimator of b. This leads to the confidence interval $[\underline{\theta},\bar{\theta}]$ for θ, where $\underline{\theta}$ and $\bar{\theta}$ are given by

$$\{\underline{\theta},\bar{\theta}\} = \frac{\bar{S}_J + \gamma^2 \hat{b} \pm \gamma\sqrt{(\bar{S}_J+\hat{b})^2 + \hat{b}^2(1-\gamma^2)(1+\lambda_J)}}{1 - \gamma^2} \quad . \tag{2.26}$$

Let $\bar{I} = [\underline{x}_q,\bar{x}_q]$ denote the corresponding confidence interval for x_q determined by $\underline{x}_q = X_{(m)} + L\,\underline{\theta}$ and $\bar{x}_q = X_{(m)} + L\,\bar{\theta}$. This is called the quadratic tail confidence interval procedure. It is location- and scale-invariant. That is, if $X_{(1)},\ldots,X_{(m)}$ are replaced by $\tau + d\,X_{(1)},\ldots,\tau + d\,X_{(m)}$, where $d > 0$, then $\underline{x}_q$ and $\bar{x}_q$ are replaced by $\tau + d\,\underline{x}_q$ and $\tau + d\,\bar{x}_q$. [A modification to this procedure in which λ_J is replaced by zero in Eq. (2.26) will be discussed in Subsection 2.4.2.]

2.4 MONTE CARLO EXPERIMENT

A Monte Carlo experiment was designed and run to compare the performance of the confidence interval procedures described in Subsections 2.1 through 2.3.

2.4.1 Experimental Design

The experimental design is a modification of the one used in Breiman, Stone and Gins [2]. Twenty underlying distribution functions were considered. They are conveniently defined in terms of four groups, each having five distribution functions. Let F_{ij} be the j^{th} distribution function in the i^{th} group. They are determined by means of a common prescription. Given i, a distribution function F_i of a positive random variable X_i is chosen as are five positive constants $b_j = b_{ij}$, $1 \leq j \leq 5$. Then F_{ij} is defined to be the distribution function of the random variable X_i^{1/b_j}. The distribution function F_i and five constants in the various groups are determined as follows:

1) (Weibull) F_1 is the standard exponential distribution function defined by $F_1(x) = 1 - e^{-x}$ for $x > 0$ and $F_1(x) = 0$ for $x \leq 0$, while $b_1 = .5$, $b_2 = .75$, $b_3 = 1$, $b_4 = 1.5$ and $b_5 = 2$. [Note that $F_{ij}(x) = 1 - e^{-x^{b_j}}$ for $x > 0$ and $1 \leq j \leq 5$.]

2) (Mixed Weibull) F_2 is defined by $F_2(x) = [F_1(x) + F_1(x/5)]/2$, while $b_1 = .6$, $b_2 = .84$, $b_3 = 1.04$, $b_4 = 1.38$ and $b_5 = 1.65$.

3) (Lognormal) F_3 is defined by $F_3(x) = \Phi(\log x)$ for $x > 0$ and $F_3(x) = 0$ for $x \leq 0$, where Φ is the standard normal distribution function, while $b_1 = .81$, $b_2 = 1.37$, $b_3 = 2.11$, $b_4 = 4.56$ and $b_5 = 10.81$.

4) (Mixed lognormal) F_4 is defined by $F_4(x) = [F_3(x) + F_3(x/5)]/2$, while $b_1 = .88$, $b_2 = 1.41$, $b_3 = 2.01$, $b_4 = 3.52$ and $b_5 = 5.60$.

To explain the choice of the constants b_{ij}, let the tail heaviness of a distribution function F having upper .1-quantile $x_{.1}$ be defined as $-\ell''(x_{.1})/[\ell'(x_{.1})]^2$, where $\ell(x) = -\log[1-F(x)]$. As pointed out in Reference 2, the tail heaviness is a reasonable measure of the departure of the upper tail of F from exponential form. It equals zero for distribution functions belonging to the two-parameter exponential model and, roughly speaking, is positive for distribution functions having heavier (i.e., more slowly decreasing) upper tails and negative for distribution functions having lighter tails. The heaviness of the Weibull distribution functions $F_{11},\ldots,F_{15}$ defining the first group are respectively .43, .14, 0, -.14 and -.22. The constants b_{1j}, $1 \leq j \leq 5$, were chosen so that the five Weibull distributions provide realistic approximations to a variety of data that have arisen in a number of air pollution studies. The constants b_{ij}, $2 \leq i \leq 4$ and $1 \leq j \leq 5$, were chosen so that F_{ij} has the same tail heaviness as F_{1j}.

The sample size n took on the values 100, 200, and 400. Given n and the underlying distribution function $F = F_{ij}$, purported 50% and 90% confidence interval procedures $\bar{I}$ for $x_{1/n}$ were compared with respect to $P_F(x_q < \bar{I})$, $P_F(x_q > \bar{I})$, $P_F(x_q \notin \bar{I})$ and $E_F(|I|/x_q)$. These quantities were estimated by averaging over 600 replications.

2.4.2 Results

Two-parameter exponential, two-parameter Weibull and quadratic tail confidence interval procedures were compared for various values of m selected more or less by trial and error (only some of which will be presented). Since

the results for the purported 50% confidence interval procedures were qualitatively so similar to those for the corresponding 90% confidence interval procedures, only the results for the latter procedures will be discussed.

Although the performance of the various procedures depends somewhat on which of the four families the underlying distribution function F_{ij} belongs to (indexed by i), it mainly depends on the tail heaviness of the distribution function, indexed by j. For this reason and for simplicity the results are presented after being aggregated by averaging over the four distribution functions having each given tail heaviness (i.e., by averaging over the four values of i for each j). An overall aggregation, obtained by averaging the results over all twenty underlying distribution functions, is also presented. See Tables 1 through 3.

Table 1 summarizes the results for n = 100. In column one the confidence procedure being used is described. In column two the tail heaviness is indicated by noting the value of the shape parameter b_{ij} which yields a Weibull distribution having the given tail heaviness. Thus the shape parameters .5, .75, 1, 1.5 and 2, respectively, correspond to the values of .43, .14, 0, -.14 and -.22 for tail heaviness. The overall average is indicated by AVG. Column three (% L is short for % Left) shows the indicated average of

$$P_{F_{ij}}(x_q < \overline{I}) \times 100\%$$

rounded off to the nearest integer (for simplicity and to provide a realistic indication of accuracy). Similarly column four (% R is short for % Right) shows the indicated average of

$$P_{F_{ij}}(x_q > \overline{I}) \times 100\% \quad ,$$

Purported 90% Confidence Intervals for $x_{1/n}$

TABLE 1. SUMMARY STATISTICS FOR n = 100

Procedure	Shape	%L	%R	%	Length
W(15)	.50	6	7	13	168
	.75	6	7	13	94
	1.00	6	7	13	65
	1.50	6	7	13	40
	2.00	6	7	13	28
	AVG	6	7	13	79
Q(40)	.50	4	9	12	125
	.75	3	7	10	96
	1.00	2	8	10	75
	1.50	3	9	11	51
	2.00	3	9	12	38
	AVG	3	8	11	77
E(15)	.50	7	19	27	77
	.75	6	9	14	67
	1.00	5	5	10	57
	1.50	5	2	8	43
	2.00	6	2	7	33
	AVG	6	7	13	55
E(10)	.50	10	12	22	101
	.75	7	7	14	80
	1.00	5	5	10	65
	1.50	4	3	7	46
	2.00	4	3	7	35
	AVG	6	6	12	66
E(15; 95%)	.50	5	13	18	95
	.75	3	5	8	82
	1.00	3	2	5	71
	1.50	2	1	4	53
	2.00	2	1	3	41
	AVG	3	5	8	68

Purported 90% Confidence Intervals for $x_{1/n}$

TABLE 2. SUMMARY STATISTICS FOR n = 200

Procedure	Shape	%L	%R	%	Length
W(20)	.50	5	7	12	125
	.75	5	7	12	74
	1.00	5	7	12	52
	1.50	5	7	12	32
	2.00	5	7	12	23
	AVG	5	7	12	61
Q(60)	.50	4	7	12	104
	.75	3	6	9	80
	1.00	3	7	10	62
	1.50	3	9	12	42
	2.00	3	10	13	30
	AVG	3	8	11	64
E(15)	.50	7	16	23	73
	.75	8	8	14	61
	1.00	5	5	11	51
	1.50	5	3	8	37
	2.00	6	2	8	28
	AVG	6	7	13	50
E(10)	.50	8	11	20	94
	.75	6	6	12	72
	1.00	6	4	10	58
	1.50	5	3	8	40
	2.00	4	3	7	30
	AVG	6	5	11	59
E(15; 95%)	.50	5	11	15	90
	.75	3	4	8	75
	1.00	3	3	5	63
	1.50	2	1	4	45
	2.00	2	1	3	34
	AVG	3	4	7	61

Purported 90% Confidence Intervals for $x_{1/n}$

TABLE 3. SUMMARY STATISTICS FOR n = 400

Procedure	Shape	%L	%R	%	Length
W(25)	.50	4	8	12	99
	.75	4	8	12	60
	1.00	4	8	12	43
	1.50	4	8	12	26
	2.00	4	8	12	19
	AVG	4	8	12	49
Q(80)	.50	3	9	12	90
	.75	3	7	10	69
	1.00	3	8	11	54
	1.50	3	9	12	36
	2.00	2	11	13	26
	AVG	3	9	12	55
E(15)	.50	6	16	22	67
	.75	5	8	13	55
	1.00	4	6	10	45
	1.50	4	4	8	32
	2.00	4	3	8	24
	AVG	5	7	12	44
E(10)	.50	8	11	19	86
	.75	5	7	13	65
	1.00	5	6	10	52
	1.50	4	4	8	35
	2.00	4	3	7	26
	AVG	5	6	11	53
E(15; 95%)	.50	4	10	14	83
	75	3	5	8	68
	1.00	2	3	6	55
	1.50	2	2	4	39
	2.00	2	1	3	29
	AVG	3	4	7	55

and column five shows the indicated average of

$$P_{F_{ij}}(x_q \notin \bar{I}) \times 100\% \quad ,$$

the numbers again being rounded off to the nearest integer. Finally, column five shows the indicated average of

$$E_{F_{ij}}(|I|/x_q) \times 100\%$$

rounded off to the nearest integer.

Let $E\big(m;100(1-\alpha)\%\big)$ denote the (purported) $100(1-\alpha)\%$ two-parameter exponential confidence interval procedure based on $X_{(1)},\ldots,X_{(m)}$ and let $W\big(m;100(1-\alpha)\%\big)$ denote the analogous Weibull procedure. It was discovered empirically that the purported $100(1-\alpha)\%$ quadratic tail confidence interval procedure given by Eq. (2.26) yields coverage percentages typically greater than $100(1-\alpha)\%$ and hence to unnecessarily long intervals. To correct this defect and to simplify the resulting procedure a modified quadratic tail procedure was employed in which λ_J is replaced by zero in Eq. (2.26). This procedure is denoted by $Q\big(m;100(1-\alpha)\%\big)$. Set $E(m) = E(m;90\%)$, $W(m) = W(m;90\%)$ and $Q(m) = Q(m;90\%)$.

The results for W(15) in the columns of Table 1 headed % L, % R and % are identical in the various rows because of the power invariance of the Weibull procedure. The average coverage percentage of W(15) is 87%. This suggests replacing W(15) = W(15;90%) by say W(15;92%) in order to obtain average coverage percentages of 90%. The modification would cause a small increase in the average length of the confidence interval.

The average coverage percentage of Q(40) is very close to 90%, but the average percentage of time the true value lies to the right of the interval is 8%, which is significantly larger than the desired value of 5%. This suggests that a better modification to Eq. (2.26) than replacing λ_J by zero might be to keep λ_J in Eq. (2.26) but adjust γ separately for $\underline{\theta}$ and $\overline{\theta}$ so that the average percentage of time the true value lies to the left of the interval and to the right of the interval both equal 5%. This modification would undoubtedly cause some increase in the average length of the confidence since the right end-point of $\overline{I}$ is more sensitive to changes in the confidence level than the left end-point.

The suggested modifications to W(15) and Q(40) would presumably lead to procedures having similar behavior. Since W(15) requires substantially more computations to implement, Q(40) appears to be the preferred procedure.

The average coverage percentage of E(15) is the same as that of W(15), namely 87%, so E(15;92%) should yield average coverage probabilities of very close to 90%. A more serious defect is that for the heaviest tailed distribution functions, the true value lies to the right of the E(15) confidence interval 19% of the time. On the other hand, the average length of E(15) is substantially less than that of Q(40). This suggests modifying E(15) to improve its average coverage percentage for the heaviest tailed distributions at the expense of increased average length. The results for two such modifications, E(10) and E(15;95%) are shown in Table 1. Clearly E(15;95%) is the better of these two procedures. It is also clear that still better modifications to E(15) could be obtained by keeping M = 15, keeping

the left end-point of the interval more or less unchanged, and increasing the right end-point [by setting $\bar{x}_q = X_{(m)} + z\,\hat{a}$ for some $z > z_{.975}$].

A similar analysis can be made of Table 2 and Table 3 for n = 200 and n = 400, respectively. The details are left to the reader.

2.5 CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDY

The results of the Monte Carlo experiment clearly indicate that the task of obtaining robust confidence intervals for the extreme quantile $x_{1/n}$ is feasible and that the two-parameter exponential and quadratic tail procedures are promising and deserve further study. But no definitive statement can yet be made that any particular procedure is best.

One attractive procedure that has not been tried out is to 1) use the upper (say) [n/2] order statistics adaptively to choose a positive number α such that the empirical distribution of $X^{\alpha}_{(1)} - X^{\alpha}_{([n/2])}, \ldots, X^{\alpha}_{([n/2]-1)} - X^{\alpha}_{([n/2])}$ is "close" to being exponential; 2) apply a (possibly modified version of) E(m) to the transformed data $X^{\alpha}_{(1)}, \ldots, X^{\alpha}_{(m)}$ obtaining an interval $[\underline{x}_q, \bar{x}_q]$; and finally, 3) apply the inverse transformation to obtain the confidence interval $[\underline{x}_q^{1/\alpha}, \bar{x}_q^{1/\alpha}]$.

A similar procedure for obtaining point estimators of x_q was suggested in Reference 2. Surprisingly, when it was tried out, the optimal value of m in the sense of mean squared error turned out to be m = n/2. A smaller value of m is probably "best" for the confidence interval problem. Indeed, it has gradually become clear that the confidence interval problem differs from the point estimation problem in one important respect that is not readily

apparent--namely that controlling bias is much more important for confidence interval procedures than for point estimators with squared error loss.

To see why this is so in the simplest possible setting, suppose that an estimator $\hat{\theta}$ of θ is normally distributed with mean $\theta + \beta$ and known variance σ^2, where the bias β satisfies $|\beta| \leq b$ for some known number b. Then the maximum possible mean square error of $\hat{\theta}$ is $\sigma^2 + b^2$. Let $0 < \alpha < 1$ and let $z_{\alpha/2}$ be defined so that

$$\int_{z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \frac{\alpha}{2} \quad .$$

Consider the confidence interval $\bar{I} = [\hat{\theta} + \tau_1, \hat{\theta} + \tau_2]$, where τ_1 is chosen as large as possible and τ_2 is chosen as small as possible subject to the constraints that $P(\theta < \hat{\theta} + \tau_1) \leq \alpha/2$ and $P(\theta > \hat{\theta} + \tau_2) \leq \alpha/2$ regardless of $\beta \varepsilon [-b,b]$. Then $\tau_1 = -z_{\alpha/2}\,\sigma - b$ and $\tau_2 = z_{\alpha/2}\,\sigma + b$ so $|\bar{I}| = 2(z_{\alpha/2}\,\sigma + b)$. For simplicity let α be chosen so that $z_{\alpha/2} = 1$. Then $|\bar{I}| = 2(\sigma + b)$. For a numerical example let $\sigma = 1$ and $b = .5$. With respect to the mean square error $\sigma^2 + b^2$, $\hat{\theta}$ is exactly as good as an unbiased estimator having standard deviation $\sqrt{1.25} = 1.12$, but with respect to the length $2(\sigma + b)$ of the corresponding confidence interval procedure, the unbiased estimator is much better.

3. QUADRATIC TAIL APPROXIMATION

3.1 THE QUADRATIC TAIL FIT

Given a distribution function $F(x)$, the general tail fitting model for $x \geq X_{(m)}$, where $X_{(m)}$ is the m^{th} highest order statistics, starts with writing $1 - F(x)$ as

$$1 - F(x) = [1 - F(X_{(m)})]\, e^{-\ell(x) - \ell(X_{(m)})} ,$$

assuming some parametric form for $\ell(x) - \ell(X_{(m)})$, and then using this fit to estimate extreme values. In general, a convenient form for defining a tail fit model is to write

$$\ell(x) - \ell(X_{(m)}) = y , \quad x \geq X_{(m)} \qquad (3.0)$$

then solve to get

$$x - X_{(m)} = L(y) . \qquad (3.1)$$

The fit is more easily defined in terms of the $L(y)$ function. For instance, the exponential tail approximation is defined by taking

$$L(y) = ay$$

or, equivalently,

$$\ell(x) - \ell(X_{(m)}) = [x - X_{(m)}]/a .$$

Then assuming m large enough so that

$$1 - F(X_{(m)}) \simeq m/n = p \quad ,$$

we get the model

$$1 - F(x) = p\, e^{-[x - X_{(m)}]/a} \quad .$$

This model has been used with success (considering its simplicity) in previous work to get estimates of percentiles high up in the tails of distributions such as lognormals, Weibull's and mixtures of these. The distributions whose tails we are attempting to approximate range over the exponential class--that is, distributions such that the maximum of n readings has, asymptotically, the first extreme value distribution. However, the asymptotics generally require a sample size much larger than usually available. For instance, our interest in the problem arose in air pollution where the interest is in estimating the expected maximum or second highest maximum of 365 readings.

If one has n observations as data, say $X_1,\ldots,X_n$, then attempting to estimate the expected maximum of N readings, N>>n, requires the assumption of a parametric model that is valid far beyond the range of the data. The range of interest to us is the range in which $1 - F(x) = \alpha/n$, $0 < \alpha < 1$. This range is of interest for two reasons. First, given n observations $X_1,\ldots,X_n$, the above range is the most extreme range that is, in some sense, "within reach" of the data. Second, many practical questions are of the form, Given these n observations, how can the expected max of n observations be estimated from them?

For example, the median of the max of n observations is defined as the solution of

$$1 - F(x) = \log 2/n \quad .$$

Furthermore, much of the classical extreme value theory is based on the x value satisfying

$$1 - F(x) = 1/n \quad .$$

Then, even to be able to apply the classical theory, this point must be estimated from the data.

For distributions whose maximums are in the domain of attraction of the first extreme value distribution, it is known that

$$\lim_{y \to \infty} P(X>x+y|X>y) \to e^{-x/a} \quad , \; x \geq 0 \quad .$$

This result gives some theoretical justification for the exponential tail approximation. But, as we mentioned above, a substantial sample size is required for the tail of the observed observations to closely follow a conditional exponential distribution. Further discussion of this issue will be given when the concept of tail heaviness is introduced (Subsection 3.3).

In order for the tail exponential to be a reasonable approximation the tail of the distribution cannot be either too "heavy" or too "light". For example, a useful intuitive notion of the range of usefulness is that the approximation works reasonably for Weibull distributions

$$1 - F(x) = e^{-(x/\beta)^{\alpha}}$$

for α in the range [1/2,2]. Put another way, the "curvature" of the tail [corresponding in some way to $\ell''(x)$] cannot be too far from constant over the range for which the tail approximation is to be used.

In this project, a search has been made for models that will give a second order approximation to the tail shape, where the exponential tail fit is considered the first order approximation. A useful model is a quadratic tail fit defined as follows: As before, let

$$\ell(x) - \ell(X_{(m)}) = y \ , \quad x > X_{(m)}$$

$$x - X_{(m)} = L(y) \quad .$$

But now take

$$L(y) = ay + \frac{b}{2} y^2 \ , \quad y > 0 \quad . \tag{3.2}$$

Our work with this model has shown that in certain areas it produces significant results in the tail estimation problem. The following subsections give a discussion of: 1) how the distribution of properties of the model are computed; 2) the concept of "tail heaviness" and estimates of the tail estimates; 3) the use of the model to estimate the expected maximum and quantiles in the range specified above; and 4) the use of the model to derive confidence intervals for parameters such as the extreme quantiles and the expected maximum.

3.2 DISTRIBUTIONAL PROPERTIES OF QUADRATIC TAIL ESTIMATES AND AN APPROXIMATION

For any tail fitting method defined by Eqs. (3.0) and (3.1), the distributional properties are simplified by the following observation:

Proposition 1. Suppose Eqs. (3.0) and (3.1) hold exactly. Let $X_{(1)} \geq \cdots \geq X_{(m)}$ be the m highest order statistics. Let $E_{(1)} \geq \cdots \geq E_{(m-1)}$ be the order statistics from a sample of size m-1 from an exponential distribution. Then the joint distribution of

$$X_{(k)} - X_{(m)}, \quad k = 1,\ldots,m-1$$

is the same as that of the variables $L(E_{(k)})$, $k = 1,\ldots,m-1$.

Proof. The variables $F(X_{(m)}) = U_{(m)}$ have the distribution of uniform order statistics. The model assumptions lead to

$$1 - U_{(k)} = [1 - U_{(m)}]\, e^{-[\ell(X_{(k)}) - \ell(X_{(m)})]}$$

or

$$-\log(1-U_k) + \log[1-U_{(m)}] = \ell(X_{(k)}) - \ell(X_{(m)}) \quad .$$

But for $k \geq m$, the lefthand side of above has the joint distribution of $E_{(1)},\ldots,E_{(m-1)}$. Hence, writing,

$$E_{(k)} = \ell(X_{(k)}) - \ell(X_{(m)})$$

and using Eq. (3.1) gives

$$X_{(k)} - X_{(m)} = L(E_{(k)}) \quad .$$

The derivation of the distributional properties of the various statistics based on the quadratic model is gotten by using the fact that, in consequence of proposition 1, the joint distribution of

$$X_{(k)} - X_{(m)} \ , \ k = 1,\ldots,m-1$$

is equal to that of

$$a\, E_{(k)} + \frac{b}{2}\,[E_{(k)}]^2 \ , \ k = 1,\ldots,m-1 \quad . \qquad (3.3)$$

Suppose, now, for example, that we want to estimate the quantile $x^*_{1/n}$ defined by

$$1 - F\left(x^*_{1/n}\right) = \frac{1}{n} \quad .$$

Write

$$1 - F\left(x^*_{1/n}\right) = [1 - F(X_{(m)})]e^{-\left[\ell\left(x^*_{1/n}\right) - \ell(X_{(m)})\right]}$$

so then, using $1 - F(X_{(m)}) \simeq m/n$ we get

$$\log m \simeq \ell\left(x^*_{1/n}\right) - \ell(X_{(m)}) \quad .$$

Therefore, using Eq. (3.0), Eq. (3.1) again is

$$x^*_{1/n} - X_{(m)} = L(\log m) = \log m + b(\log m)^2/2$$

resulting in the estimate

$$x^{*}_{1/n} = X_{(m)} + \log m \left[a + \left(\frac{\log m}{2}\right) b\right] \quad .$$

Thus, the problem becomes to find a good estimate for the parameter

$$a + \left(\frac{\log m}{2}\right) b \quad .$$

Similarly, estimating the expected max, the second highest maximum, etc., can all be formulated in terms of finding estimates for a parameter of the form

$$\theta = La + Mb \tag{3.4}$$

where L,M are known. Since $X_{(j)}$ is a linear expression in a,b then it is natural to look for estimates of θ that are linear in $X_{(j)}$, $j \leq m$. It is convenient to write these estimates as

$$\hat{\theta} = \sum_{1}^{m-1} k\, \omega_k [X_{(k)} - X_{(k+1)}] \quad . \tag{3.5}$$

Estimates of this type can be easily computed from the data once the values of ω_k are given.

In Appendix I, the mean and variance of the estimate of Eq. (3.5) are derived. The main results are

$$E\hat{\theta} = \left(\sum_{1}^{m-1} \omega_k\right) a + \left(\sum_{1}^{m-1} \mu_k\, \omega_k\right) b \tag{3.6}$$

where

$$\mu_k = \sum_{\ell=k}^{m-1} (1/\ell) \quad . \tag{3.7}$$

Thus, the conditions for an estimate to be unbiased are

$$L = \sum_{1}^{m-1} \omega_k \ , \quad M = \sum_{1}^{m-1} \mu_k \omega_k \quad . \tag{3.8}$$

The variance of $\hat{\theta}$ is given as follows: Define

$$h = (b/a)$$

$$\gamma_k = \frac{1}{k} \sum_{\ell=1}^{k} \omega_k \tag{3.9}$$

$$\mu_k^{(2)} = \sum_{k}^{m-1} 1/\ell^2 \quad ;$$

then

$$\mathrm{Var}(\hat{\theta})/a^2 = \sum_{1}^{m-1} (\omega_k + h\ \mu_k \omega_k + h\gamma_k)^2 + h^2 \sum_{1}^{m-1} \mu_k^{(2)}\ \omega_k^2 \quad . \tag{3.10}$$

Note, incidentally, that for

$$\omega_k = 1/m\text{-}1 \ ,$$

$$\hat{\theta} = \frac{1}{m\text{-}1} \sum_{1}^{m-1} [X_{(j)} - X_{(m)}]$$

which is the estimate of the mean in the exponential tail approximation.

Using Eq. (3.10), minimum variance unbiased estimators can be derived as well as estimators that minimize the least squares loss for given values of $h = a/b$. The details will be discussed in Appendix II.

However, minimizing Eq. (3.10) even with a small approximation as done in Appendix II [the last term of Eq. (3.10) is discarded] leads to calculations which necessitate a small computer or hand programmable calculator. For this reason we looked at an approximation which is applicable not only to quadratic $L(y)$, but general $L(y)$.

Write

$$\hat{\theta} = \sum_{1}^{m-1}{}_{k} \omega_k [L(E_k) - L(E_{k+1})] \quad .$$

Using a first order Taylor expansion

$$L(E_{(k)}) - L(E_{(k+1)}) = [E_{(k)} - E_{(k+1)}] L'\left(E_{(k)} + \theta[E_{(k)} - E_{(k+1)}]\right) \quad .$$

If $L'(y)$ is slowly changing, then the approximation

$$L'\left(E_{(k)} + \theta(E_k - E_{k+1})\right) \simeq L'\left(E(E_{(k)})\right) = L'(\mu_k)$$

seems reasonable. Now

$$E_{(k)} - E_{(k+1)} = Z_k/k \ , \ k = 1,\ldots,m-1$$

where $Z_1,\ldots,Z_{m-1}$ are independent unit exponentials. Thus

$$\hat{\theta} \simeq \sum_{1}^{m-1} \omega_k Z_k L'(\mu_k) \quad .$$

Then the variance of $\hat{\theta}$ is

$$\mathrm{Var}(\hat{\theta}) = \sum_{1}^{m-1} \omega_k^2 [L'(\mu_k)]^2 \quad .$$

The expectation is

$$E(\hat{\theta}) = \sum_{1}^{m-1} \omega_k \, L'(\mu_k) \quad .$$

Suppose that $L(y) = L(y,\underline{\beta})$ depends on some multidimensional parameter $\underline{\beta}$, and suppose that $\hat{\theta}$ is an estimate of $\theta(\underline{\beta})$. Then θ is approximately unbiased if

$$\theta(\underline{\beta}) = \sum \omega_k \, L'(\mu_k,\underline{\beta}) \quad .$$

To get the minimum LSE estimate at $\underline{\beta} = \underline{\beta}_0$, look at the square error

$$LSE = Var(\hat{\theta}) + B^2$$

where

$$B = \theta(\underline{\beta}_0) - \sum \omega_k \, L'(\mu_k,\underline{\beta}_0) \quad .$$

To minimize the LSE, take its derivative with respect to ω_k, getting

$$\omega_k \left[L'(\mu_k,\underline{\beta}_0) \right]^2 - B \, L'(\mu_k,\underline{\beta}_0) = 0 \quad ;$$

so

$$\omega_k = B/L'(\mu_k,\underline{\beta}_0) \quad .$$

Not unexpectedly, if $L'(y)$ is increasing, the coefficients ω_k decrease and conversely. For $L(y) = ay$, the coefficients ω_k are constant, giving the exponential tail approximation. B can be evaluated using

$$B = \theta(\beta_0) - \sum \omega_k L'(\mu_k, \underline{\beta}_0)$$

$$= \theta(\underline{\beta}_0) - (m-1)\, B \quad ;$$

so

$$B = \theta(\underline{\beta}_0)/m \quad .$$

If we want to get an unbiased minimum variance estimator of $\underline{\beta}_0$, then we minimize $\mathrm{Var}(\hat{\theta})$ subject to $\theta(\underline{\beta}_0) = \sum \omega_k L'(\mu_k, \beta_0)$. This gives the same solution as above.

Something different can be done with the quadratic tail approximation to produce estimators that are unbiased over the entire range of a,b for which the tail fit is reasonable. The LSE minimum estimator of $\theta_0 = L\, a_0 + M\, b_0$ is given by

$$\omega_k = \frac{L\, a_0 + M\, b_0}{a_0 + b_0\, \mu_k} = \frac{L + M\, h_0}{1 + h_0\, \mu_k}$$

and will give low LSE only in a vicinity of $h = h_0$. The unbiased requirement, however, can be put in terms of a and b separately because of the linearity in a,b. That is, we can write

$$\theta = La + Mb$$

$$E\hat{\theta} = a \sum \omega_k + b \sum \omega_k \mu_k$$

and require that the ω_k satisfy both

$$L = \sum \omega_k$$

and

$$M = \sum \omega_k \mu_k \quad .$$

Minimizing the variance under these two restrictions at h_0 gives

$$\omega_k = \frac{\lambda_0 + \lambda_1 \mu_k}{(1 + h_0 \mu_k)^2}$$

where λ_0 and λ_1 are determined by the constraints.

However, every estimator we have simulated which satisfies the two above constraints, no matter how complicated or simple it is, has had small bias for all distributions tested except the heaviest tailed lognormal.

We give some evidence in Appendix II that the approximate solution gives results very close to those of the exact (well, almost!) solution.

3.3 TAIL HEAVINESS ESTIMATES

In the previous report on work in this area, a definition of tail heaviness was proposed as meeting certain reasonable requirements (Reference 1, pp. 18-20). The characterization did not consist of a single number, but was a

local measure of the curvature of the distribution. The tail heaviness H(p) at p, $0 \leq p \leq 1$ was defined as follows: Let

$$\gamma(x) = \ell''(x)/[\ell'(x)]^2 , \quad x > 0$$

and set

$$H(p) = -\gamma(x_p)$$

where x_p is the p^{th} quantile of the distribution. For exponential distributions, H(p) = 0. A positive value of H(p) corresponds to a tail that is "heavy" relative to the exponential, and a negative H(p) to a tail lighter than the exponential. For the Weibull and lognormal distributions studied in Reference 1, the tail heaviness is tabled below.

Weibull

		b						
		.5	.75	1.0	1.25	1.5	1.75	2.0
	.3	.83	.27	0	-.17	-.27	-.36	-.42
	.2	.62	.21	0	-.12	-.21	-.27	-.31
	.1	.43	.14	0	-.09	-.14	-.19	-.22
p	.05	.33	.11	0	-.07	-.11	-.14	-.17
	.01	.22	.07	0	-.04	-.07	-.09	-.11
	.005	.19	.06	0	-.04	-.06	-.08	-.09
	.001	.15	.05	0	-.03	-.05	-.06	-.07

Lognormal

		b						
		.5	.75	1.0	1.25	1.5	1.75	2.0
	.3	1.17	.60	.31	.14	.02	-.06	-.12
	.2	1.03	.55	.31	.17	.08	.01	-.04
	.1	.68	.36	.20	.11	.04	.00	-.03
p	.05	.61	.34	.20	.12	.07	.03	.00
	.01	.52	.29	.19	.13	.09	.06	.03
	.005	.46	.27	.17	.12	.08	.05	.03
	.001	.39	.23	.15	.10	.06	.04	.02

For the Weibull's with b = .75 to 1.5, and for the lognormals with $b \geq 1.25$ there is a minimal curvature problem for $p \leq .1$. The curvature problem is more severe for the extremely heavy and light tailed Weibull's and the heavy lognormals. Since the asymptotic extreme value theory depends on the condition $H(p) \to 0$, the size of H(p) for these latter distributions for p as small as .001 indicates the inapplicability of the theory for fairly large sample sizes.

Using the quadratic model we have:

Proposition 1. The tail heaviness at $p = 1 - F(X_{(m)})$ is equal to

$$h = b/a \quad .$$

Proof: Since $\ell'(x) - \ell(X_{(k)}) = y$, then

$$\ell'(x) = \frac{dy}{dx} = 1\Big/\left(\frac{dx}{dy}\right)$$

$$\ell''(x) = -\frac{d}{dy}\left(\frac{dy}{dx}\right) \cdot \frac{dy}{dx}$$

$$= -\left(\frac{d^2x}{dy^2}\right)\Big/\left(\frac{dx}{dy}\right)^3 \quad .$$

Using

$$x - X_{(k)} = ay + \frac{b}{2}y^2$$

gives

$$\frac{dx}{dy} = a + by$$

$$\frac{d^2x}{dy^2} = b \quad .$$

Therefore, at $x = X_{(m)}$ or $y = 0$,

$$-\gamma(X_{(m)}) = b/a = h \quad .$$

Therefore, the quadratic model can be used to get estimates of the tail heaviness. For a simulation we constructed four estimators. The first three were of the following type: Denote

$$\overline{S}_k = \frac{1}{k-1}\sum_{1}^{k-1}[X_{(j)} - X_{(k)}] \quad .$$

Then we used a split half approach and look for a linear combination

$$C\,\bar{S}_{[m/2]} + D\,\bar{S}_m$$

which will give an unbiased estimate of b. Using conditions in Eq. (3/8) gave the result

$$C = -D = 1/\mu_{[m/2]} \quad .$$

The selection of the denominator posed a problem. Using an unbiased estimator for a gave a noisy estimate for h. Finally, we decided to use $\bar{S}_m$ as the denominator. For the last estimator we used an unbiased estimator at b whose variance was a minimum at h = .25.

For the simulation, we used a sample size of 200 with the Weibulls and lognormals as above. Three different values of m were selected, m = 20, 30, 60. The first three estimators, then, had the form

$$\hat{h}_m = \frac{1}{\mu_{[m/2]}}\left(\frac{\bar{S}_{[m/2]}}{\bar{S}_m} - 1\right), \; m = 20,\ 30,\ 60 \quad .$$

For the last estimator m = 60 was used, and it is denoted by $\hat{\hat{h}}_{60}$.

Altogether, 1000 runs were made, each run generating the top 60 order statistics of the 14 distributions using Marsaglia's "Super-Duper" uniform random number generator, and using the inverse transformation to get the order statistics of the distributions desired. The results are tabled below, giving the average and the standard deviations of the estimates.

Weibull

		b = .5	.75	1.0	1.25	1.5	1.75	2.0
$\hat{h}_{20}$	AV	.258	.083	-.003	-.053	-.086	-.109	-.126
	SD	.329	.315	.307	.303	.300	.298	.296
$\hat{h}_{30}$	AV	.314	.105	.003	-.057	-.097	-.124	-.145
	SD	.275	.262	.255	.250	.248	.246	.244
$\hat{h}_{60}$	AV	.422	.134	-.008	-.092	-.146	-.184	-.212
	SD	.199	.188	.181	.177	.174	.171	.170
$\hat{\hat{h}}_{60}$	AV	.427	.134	-.006	-.087	-.140	-.177	-.204
	SD	.187	.159	.148	.142	.139	.137	.136

Lognormal

		b = .5	.75	1.0	1.25	1.5	1.75	2.0
$\hat{h}_{20}$	AV	.525	.296	.180	.111	.066	.033	.010
	SD	.377	.357	.344	.335	.330	.327	.324
$\hat{h}_{30}$	AV	.592	.331	.198	.119	.068	.031	.004
	SD	.309	.290	.277	.269	.264	.260	.258
$\hat{h}_{60}$	AV	.693	.371	.205	.107	.043	-.002	-.035
	SD	.230	.217	.207	.201	.197	.194	.192
$\hat{\hat{h}}_{60}$	AV	.745	.388	.214	.113	.049	.004	-.030
	SD	.273	.218	.190	.177	.169	.164	.161

Since the denominator is an unbiased estimate of a + b, instead of a, the estimates tend to reflect the values of H(p) at values of p larger than m/n. For all 14 distributions, the estimates $\hat{h}_{60}$ and $\hat{\hat{h}}_{60}$ have average values close to H(.1). The averages of $\hat{h}_{30}$ are close to H(.05), and the averages of $\hat{h}_{20}$ are slightly above the H(.01) values.

Of course, the smaller m is, the higher the SD's of the estimates. In use, either $\hat{h}_{60}$ or $\hat{\hat{h}}_{60}$ is preferable in terms of variance, if p = .1 is in the tail range of interest in the problem. Note that $\hat{\hat{h}}_{60}$ almost always has a significantly lower variance than $\hat{h}_{60}$, but this has to be balanced against the difficulty of computing the coefficients.

If these estimates of tail heaviness are used to detect departure from exponentiality, then, using a 2SD rule of thumb, the true value of H(.1) would have to satisfy $|H(.1)| \geq .3$ before the departure could be reliably detected. Thus, only the curvature of the heaviest tailed Weibull and the two heaviest-tailed lognormals can usually be detected. However, these are the distributions that cause the largest absolute errors in the exponential tail estimates.

3.4 ESTIMATES OF THE EXPECTED MAXIMUM

As a test of the quadratic model and exponential tail estimators, a simulation study was designed to estimate the bias and variance of a variety of tail estimators of the expected maximum of the 14 Weibull and lognormal distributions.

Twelve estimators were computed and compared, again using a sample size of 200, 1000 runs, the Marsaglia random number generator and inverse functions.

The twelve estimators were in 4 groups: 1) the maximum $X_{(1)}$ was used as an estimate for $E\ X_{(1)}$; 2) three exponential tail estimates corresponding to m = 20, 30, 60; 3) four quadratic minimum variance unbiased tail estimates with the variance minimized respectively at

$h = .5, .25, 0.0, -.15$ using $m = 60$; 4) four quadratic minimum squared error estimates with the minimization carried out at $h = .5, .25, 0.0, -.15$ and using $m = 60$.

The quadratic model was used to construct the third and fourth groups of estimates. The derivation is based on the following: Write

$$X_{(1)} = X_{(1)} - X_{(m)} + X_{(m)} \quad .$$

Now by the model

$$X_{(1)} - X_{(m)} = a\, E_{(1)} + \frac{b}{2} E^2_{(1)} \quad .$$

Use the fact that

$$E(E_{(1)}) = \mu_1 = \sum_{1}^{m-1} 1/\mu$$

and

$$E(E^2_{(1)}) = (\mu_1)^2 + \mu_1^{(2)} \quad .$$

For $m = 60$, $\mu_1 = 4.655$, $\mu_1^{(2)} = 1.645$, so,

$$E(X_{(1)} - X_{(m)}) = 4.655a + 11.657b \quad .$$

Therefore, estimating the parameter

$$\hat{\theta} = 4.655a + 11.657b$$

by $\hat{\theta}$ gives the estimate $X_{(m)} + \hat{\theta}$ for $E\, X_{(1)}$.

Denote by RMSE the root mean squared error. The tables below give the value of E $X_{(1)}$ and the RMSE of $X_{(1)}$ as an estimate of E $X_{(1)}$.

Weibull

b	.5	.75	1.0	1.25	1.5	1.75	2.0
$EX_{(1)}$	36.2	10.7	5.9	4.1	3.2	2.7	2.4
RMSE$[X_{(1)}]$	17.2	3.2	1.3	.70	.46	.33	.25

Lognormal

$EX_{(1)}$	63.2	14.8	7.4	4.9	3.7	3.1	2.7
RMSE$[X_{(1)}]$	66.9	8.9	3.1	1.6	.98	.68	.51

The RMSE$[X_{(1)}]$ was used as a benchmark, and the ratio of the RMSE's of all other estimates to the RMSE$[X_{(1)}]$ was computed. The results are tabled below.

RATIOS OF RMSE's

Weibull

Estimates	b .5	.75	1.0	1.25	1.5	1.75	2.0
exp (20)	.61	.61	.66	.73	.80*	.86*	.92
exp (30)	.66	.57	.59	.69	.80*	.91	1.02
exp (60)	.95	.59	.49	.71	1.02	1.30	1.57
UNBMIN (h=.5)	.63	.89	1.11	1.33	1.55	1.76	1.95
UNBMIN (h=.25)	.65	.83	.98	1.12	1.26	1.40	1.53
UNBMIN (h=0.0)	.75	.83	.88	.92	.97	1.01	1.05
UNBMIN (h=-.15)	1.13	1.11	1.10	1.09	1.09	1.08	1.08
$MINE^2$ (h=.5)	.46*	1.12	2.47	3.69	4.73	5.60	6.35
$MINE^2$ (h=.25)	.64	.56*	1.39	2.23	2.96	3.56	4.10
$MINE^2$ (h=0.0)	.87	.62	.48*	.66*	.95	1.21	1.47
$MINE^2$ (h=-.15)	.43	.93	.92	.91	.90	.90	.90*
Minimum	.46	.56	.48	.66	.80	.86	.90

*Minimum value.

Lognormal

Estimates	b .5	.75	1.0	1.25	1.5	1.75	2.0
exp (20)	.48	.53	.54	.54	.55	.57	.59
exp (30)	.52	.56	.54	.51	.50	.51	.52
exp (60)	.61	.68	.61	.51	.44*	.41*	.43
UNBMIN (h=.5)	.42*	.51	.64	.75	.86	.95	1.04
UNBMIN (h=.25)	.45	.55	.64	.72	.79	.86	.91
UNBMIN (h=0.0)	.63	.69	.74	.77	.80	.82	.84
UNBMIN (h=-.15)	1.17	1.14	1.12	1.11	1.10	1.10	1.10
$MINE^2$ (h=.5)	.54	.37*	.42	88	1.37	1.81	2.21
$MINE^2$ (h=.25)	.48	.52	.35*	.43*	.67	.98	1.24
$MINE^2$ (h=0.0)	.62	.70	.64	.54	.46	.42	.42*
$MINE^2$ (h=-.15)	.92	.91	.91	.90	.90	.90	.90
Minimum	.42	.37	.35	.43	.44	.41	.42

*Minimum value.

The fundamental issue in the behavior of these estimates is the trade-off between bias and variance. The two estimators exp (30) and exp (20) have the best overall performance with average ratios of:

	Weibull	Lognormal
exp (20)	.74	.54
exp (30)	.75	.52

The estimate exp (20) is better at the extreme values of heaviness, where bias is more of a factor while exp (30) performs better for moderate to small values of H where its variance is smaller.

The other estimators performed as expected with one exception, discussed below. The unbiased estimators, as will be shown in the percent bias tables below, lived up to their billing and had very little bias for any of the distributions except the heaviest tailed lognormal. However, the payment was in terms of variance. Except near the values of h at which their variance was minimized, their variances were large and produced inflated RMSE's. The minimum squared error estimates also lived up to their billing by producing either the minimum or near minimum ratio near the values of h at which they were optimized. Their difficulty was that a bias which is small at one value of h can be large at other values. The large bias led to large ratios away from the value of h at which they were optimized.

A look at the percent of bias, that is,

$$100 \cdot \frac{E\ X(1) - \overline{est}}{E\ X(1)}$$

is revealing.

Percent Bias - Weibull

Estimator	.5	.75	1.0	b 1.25	1.50	1.75	2.0
$X_{(1)}$	-.1	-.1	-.1	-.1	-.1	.1	-.1
exp (20)	20.2	6.2	.5	-2.2	-3.5	-4.0	-4.4
exp (30)	26.1	8.4	.3	-3.6	-5.6	-6.3	-7.0
exp (60)	38.6	13.4	-.2	-7.5	-11.5	-13.4	-14.7
UNBMIN (.5)	.4	-2.9	.4	3.8	6.2	8.0	8.9
UNBMIN (.25)	.5	-2.1	.5	3.1	4.9	6.3	6.9
UNBMIN (0.0)	.5	-.9	.4	1.6	2.5	3.2	3.4
UNBMIN (-.15)	-.3	.1	-.2	-.4	-.6	-.6	-.8
$MINE^2$ (.5)	15.3	-28.3	-50.4	-60.7	-64.8	-65.5	-65.1
$MINE^2$ (.25)	27.2	-8.2	-26.6	-35.7	-39.8	-41.1	-41.5
$MINE^2$ (0.0)	39.5	14.6	1.1	-6.2	-10.2	-12.2	-13.5
$MINE^2$ (-.15)	11.4	7.4	4.6	2.9	1.8	1.2	.5

Lognormal

$X_{(1)}$	-5.3	-3.2	-2.2	-1.7	-1.4	-1.0	-.8
exp (20)	42.1	23.2	13.2	7.6	4.4	2.5	1.2
exp (30)	50.4	29.0	16.8	9.7	5.4	2.8	1.1
exp (60)	63.3	39.3	23.3	13.2	6.8	2.6	-.1
UNBMIN (.5)	25.6	8.4	3.4	2.3	2.5	3.0	3.5
UNBMIN (.25)	21.1	6.7	2.6	1.7	1.8	2.3	2.7
UNBMIN (0.0)	10.0	2.5	.6	.2	.4	.7	1.0
UNBMIN (-.15)	-10.1	-4.7	-2.8	-2.0	-1.6	-1.3	-1.0
$MINE^2$ (.5)	56.2	17.1	-8.5	-23.7	-32.3	-37.0	-39.4
$MINE^2$ (.25)	61.1	28.6	7.1	-5.9	-13.7	-18.1	-20.8
$MINE^2$ (0.0)	63.9	40.2	24.3	14.2	7.8	3.7	.9
$MINE^2$ (-.15)	7.9	7.9	6.5	5.1	4.1	3.4	2.9

Notice that the estimators certified as unbiased by the quadratic model do, indeed, have very small biases. The best is the unbiased estimate whose variance is minimized at h = 0. Its average percent bias (absolute values) is 2.0%. The only problem, again, is with the heaviest tailed log-normal where the bias rises to 10%.

To show that even simple estimates satisfying

$$L = \sum \omega_k$$

$$M = \sum \omega_k \mu_k$$

will be relatively unbiased, we selected an estimate of the form

$$\lambda_0 S_{(30)} + \lambda_1 S_{(15)} \quad .$$

The values of λ_0, λ_1 satisfying the two unbiasedness conditions are

$$\lambda_1 = (M-L)/\mu_{15}$$

$$\lambda_0 = L - \lambda_1 \quad ,$$

where $\mu_{15} = \sum_{15}^{24} 1/k$. Using this with values of L,M adjusted to estimate $E\, X_{(1)}$ with the top 30 order statistics, the estimator gave the following bias:

Percent Bias

	b						
	.5	.75	1.0	1.25	1.50	1.75	2.0
Weibull	.3	-.8	.2	1.1	1.7	2.2	2.3
Lognormal	13.2	3.4	.9	.4	.4	.6	.8

Because of the fact that the estimators that minimize the squared error are sensitive to the choice of h used, another experiment was carried out for the distributions such that $H(.1) > 0$. In this simulation, the unbiased estimator $\hat{h}$ of $h = a/b$ was first computed using the upper 60 order statistics; then, the minimum squared error estimate corresponding to the value $\max(\hat{h},0)$ was calculated. The results are promising, as given in the table below of the average ratios of the RMSE to that of $X_{(1)}$.

AVERAGE RATIOS OF RMSE/RMSE$[X_{(1)}]$ FOR $H(.1) > 0$

	Weibull	Lognormal
exp (20)	.61	.53
exp (30)	.62	.53
2 step est.	.57	.50

Although the improvement is not very large, it is suggestive for future work.

3.5 CURVATURE AND TRANSFORMED TAILS

One somewhat strange bit of behavior is given by the estimates optimized at $h = -.15$. On the one hand, their biases are consistently small, doing better for large positive H values than the unbiased estimates optimized at positive values of h. We conjecture that this is due to the fact that

minimizing the variance at a given value of h does not enhance the general unbiasedness properties. However, a harder to explain phenomenon is that the minimum squared error estimate at h = -.15 does not give a significant decrease in the squared error for the Weibull distributions with negative H. In fact, for the four Weibulls with negative H, exp (20) is a consistent improvement on $MINE^2(-.15)$ except for a slight difference at b = 2.0.

At first, we thought that this indicated some deficiency of the quadratic model for light tailed distributions and that a different method of selecting the coefficients ω_k in the estimate

$$\hat{\theta} = \sum_{1}^{m-1} k\, \omega_k \, [X_{(k)} - X_{(k+1)}]$$

would bring the ratio of MSE's down. We concluded that the difficulty was more fundamental. First we noticed that even with the estimators exp (20), exp (30), the ratios go up rapidly as the Weibulls become light tailed, i.e., b > 1. On investigating the cause, it turned out that the difficulty was not really with the bias (although that contributed) but instead with the fact that the standard deviations of exp (20), exp (30), were, surprisingly enough, not that much less than the standard deviation of $X_{(1)}$. The results are given below.

SD AND BIAS OF ESTIMATORS

WEIBULL DISTRIBUTION

		b				
		1.0	1,25	1.50	1.75	2.0
$X_{(1)}$	SD	1.38	.70	.46	.33	.25
	bias	0.00	0.00	0.00	0.00	0.00
exp (20)	SD	.84	.50	.35	.26	.21
	bias	0.00	.10	-.11	-.11	-.11
exp (30)	SD	.75	.46	.32	.24	.20
	bias	0.02	-.15	-.18	-.17	-.17

Thus, for instance, even if exp (30) were unbiased at b = 2.0, the ratios of RMSE's would be .80. This led to the conjecture that the essential difficulty was with the linear form of the estimate; that at least with light tailed distributions, one could not construct a linear combination of $X_{(k)} - X_{(k+1)}$ that was not too biased without a resulting SD close to that of $X_{(1)}$.

This led to the idea of seeing how much an optimum or nearly optimum transformation could improve the RMSE. If X has a Weibull distribution with parameter α then X^{α} has an exponential distribution. Therefore, given $X_{(m)}$, the differences

$$X^{\alpha}_{(k)} - X^{\alpha}_{(m)}, \qquad k > m$$

have exactly the distribution of exponential order statistics. That is, the transformed variables $X^{\alpha}_{(k)}$ are exactly fit by an exponential tail. Thus, the exponential tail estimate

$$\hat{e} = X_{(m)}^{\alpha} + \mu_1 \left\{ \frac{1}{m-1} \sum_{1}^{m-1} [X_{(k)}^{\alpha} - X_{(m)}^{\alpha}] \right\}$$

should give a good estimate of $E\ X_{(1)}^{\alpha}$. Then take $(\hat{e})^{1/\alpha}$ to be the estimator of $E\ X_{(1)}$. We assumed that α was known for the Weibulls. That is, for the Weibull raised to the power b, we took $\alpha = b$.

For the lognormals, it is not clear what the optimum transformation is. We selected the powers α based on a comparison of the lognormal heaviness to the corresponding Weibull heaviness. So, as a guess, we took the values of α corresponding to b to be given by

b =	.5	.75	1.0	1.25	1.50	1.75	2.0
α =	.4	.5	.6	.7	.8	.9	1.0

The results, expressed as ratios to $RMSE[X_{(1)}]$ are

Ratios

	b						
	.5	.75	1.0	1.25	1.5	1.75	2.0
Weibull	.44	.47	.49	.49	.50	.50	.51
Lognormal	.41	.40	.39	.39	.40	.41	.43

The Weibull results are a lower bound under the assumptions that only the top 60 order statistics are used and that the scale is unknown. Of course, there is the question of how much bias has been introduced by the approximation $E\ X_{(1)} \simeq [E\ X_{(1)}^{\alpha}]^{1/2}$. For the Weibulls, the bias of the estimate is 3% at b = .5 and less than .6% for the others. The lognormal

estimates have considerably larger bias at the heavier end, rising to 8% at b = 1.0, 17% at b = .75 and 39% at b = .5. The indications are that having smaller values of α at the heavier tailed lognormals would have given further reductions in the ratios.

To get a comparison of the best linear fit based on the quadratic model, we ran the approximate solution to the minimum RMSE optimized at h = .6, .4, .2, 0, -.2. The table below is the minimum ratio over all of the estimates.

RATIOS FOR "BEST" LINEAR FIT

("Best" = best quadratic model linear fit)

	b						
	.5	.75	1.0	1.25	1.50	1.75	2.0
Weibull	.50	.49	.48	.66	.95	.95	.91
Lognormal	.43	.41	.38	.38	.46	.42	.42

For all of the lognormal distributions the best linear and the transformed estimates are comparable in terms of RMSE. They are comparable for the Weibull for $b \leq 1.0$.

This gives some evidence that long and short tailed distributions necessitate different procedures to give good estimates of $E\, X_{(1)}$. For the long tailed distributions a transformation is not needed to "uncurl" the tail; the appropriate linear combination of order statistics will do almost as well as the best transformed estimate. In the short tailed cases (Weibull with

$b > 1.0$) a power transformation is essential to significantly reduce the SD of the estimates.

3.6 CONFIDENCE INTERVALS

Suppose that the parameter of concern is of the form

$$X(m) + La + Mb \quad .$$

If 100% confidence intervals are found for

$$\theta = a + (M/L)b = a + Nb$$

say $\underline{U}, \overline{U}$, that is

$$P_\theta(\underline{U} \leq \theta \leq \overline{U}) = 1 - Q \quad ,$$

then 100Q% confidence intervals for $X(m) + La + Mb$ are (approximately) given by

$$[X(m) + L\underline{U},\ X(m) + L\overline{U}] \quad .$$

Start by finding a simple unbiased estimator for θ. Take the estimator to be of the form

$$\overline{S}_J$$

where

$$\overline{S}_J = \frac{1}{J-1} \sum_{1}^{J-1} [X_{(k)} - X_{(J)}]$$

where J is the value to be determined. Since $\omega_k = 1/J-1$, $k \leq J-1$, and zero for $k \geq J$, the expectation of $\bar{S}_J$ is, by Eq. (3.8), equal to

$$a + \left(\frac{1}{J-1} \sum_{1}^{J-1} \mu_k\right) b \quad .$$

An easy calculation gives

$$\frac{1}{J-1} \sum_{1}^{J-1} \mu_k = 1 + \mu_J \quad .$$

Hence J is determined through the equation

$$\mu_J = N - 1 \quad .$$

The well-known approximation

$$\sum_{i=1}^{k} 1/i \sim \log k + \gamma$$

leads to $\mu_J \sim \log [(m-1)/(J-1)]$ so that we get the equation

$$J - 1 \simeq (m-1)\, e^{-(N-1)} \quad . \tag{3.11}$$

To compute the variance of $\bar{S}_J$, use Eq. (3.10) and the fact that

$$\gamma_k = \begin{cases} 1/J-1 \ , & k \leq J-1 \\ 1/k \ , & k \geq J \end{cases} \ ,$$

to get

$$\mathrm{Var}(\bar{S}_J) = C_1\, a^2 + C_2\, ab + C_3\, b^2$$

where

$$C_1 = \sum_{1}^{m-1} \omega_k^2 = 1/J\text{-}1$$

$$C_2 = \frac{2}{J-1} \sum_{1}^{J-1} \left(\mu_k \frac{1}{J-1} + \frac{1}{J-1} \right)$$

$$= \frac{2}{(J-1)^2} [2(J-1) + (J-1)\, \mu_J]$$

$$= \frac{2}{(J-1)} (1 + N)$$

$$C_3 = \frac{1}{(J-1)^2} \sum_{1}^{J-1} (\mu_k+1)^2 + \sum_{J}^{m-1} (1/k)^2 + \frac{1}{(J-1)^2} \sum_{1}^{m-1} \mu_k^{(2)} .$$

Now

$$\sum_{J}^{m-1} (1/k)^2 = \mu_J^{(2)}$$

$$\sum_{1}^{m-1} \mu_k^{(2)} = \sum_{1}^{m-1} \sum_{j=k}^{m-1} 1/j^2 = \mu_1$$

$$\sum_{1}^{J-1} (\mu_k+1)^2 = (J-1)[(2+\mu_J)^2 + 1] - \mu_1 ,$$

where

$$\mu_1' = \sum_{j=1}^{J-1} 1/j \quad .$$

(This last calculation is carried out by writing $\mu_k = \mu_J + \mu_k'$ where $\mu_k' = \sum_{j=k}^{J-1} 1/j$.) Hence

$$C_3 = \frac{1}{J-1} \left[(2 + \mu_J)^2 + 1\right] + \frac{\mu_J}{(J-1)^2} + \mu_J^{(2)}$$

or, assuming J, say, ≥ 10, then

$$C_3 \simeq \frac{1}{J-1} \left[(2 + \mu_J)^2 + (J-1)\, \mu_J^{(2)}\right] \quad .$$

And since $2 + \mu_J = N + 1$,

$$\mathrm{Var}(\bar{S}_J) = \frac{1}{(J-1)} \left\{a^2 + 2(N+1)ab + (N+1)^2 b^2 + b^2\left[1 + (J-1)\, \mu_J^{(2)}\right]\right\} \quad .$$

Note that $\lambda_J = (J-1)\, \mu_J^{(2)} \simeq 1 - (J-1)/(m-1)$.

We can write the above as

$$\mathrm{Var}(\bar{S}_J) = \frac{1}{(J-1)} \left\{[a + (1+N)b]^2 + (1 + \lambda_J)\, b^2\right\} \quad . \qquad (3.12)$$

The righthand side of Eq. (3.12) involves the unknown parameters a and b. However, we can write

$$a + (1+N)b = a + Nb + b = \theta + b \quad .$$

We can get an estimate of b by using an expression of the form

$$\hat{b} = C(\bar{S}_J - \bar{S}_m) \quad .$$

Looking at Eq. (3.8), $\hat{b}$ will be an unbiased estimate of b if

$$C = 1/\mu_J = 1/N\text{-}1 \quad .$$

The key to the remaining part of the computation is the assumption that $\bar{S}_J$ has an approximately normal distribution. Under this assumption

$$P_\theta\left(-z \cdot SD(\bar{S}_J) \leq \bar{S}_J - \theta \leq z\ SD(\bar{S}_J)\right) = 1 - Q \quad , \tag{3.13}$$

where $SD(\bar{S}_J)$ is the standard deviation of $\bar{S}_J$ and z is computed from the normal tables, i.e., for a unit normal Z,

$$P(Z \geq z) = Q/2 \quad .$$

Write the inequality inside the probability in Eq. (3.13) as

$$(\bar{S}_J - \theta)^2 \leq z^2\ Var(\bar{S}_J)$$

or putting $\gamma = z/(1\text{-}J)$, and using Eq. (3.12)

$$(\bar{S}_J - \theta)^2 \leq \gamma^2\ [(\theta + b)^2 + (1 + \lambda_J)\ b^2] \quad .$$

Simplifying leads to

$$\theta^2(1 - \gamma^2) - 2\theta(\bar{S}_J + \gamma^2 b) - \gamma^2\ b^2(2 + \lambda_J) + \bar{S}_J^2 \leq 0 \quad .$$

Solving gives the following expression for the roots:

$$\frac{(\bar{S}_J+\gamma^2 b) \pm \sqrt{(S_J+\gamma^2 b)^2 - S_J^2(1-\gamma^2) + (1-\gamma^2)\gamma^2 b^2(2+\lambda_J)}}{1-\gamma^2} \quad ;$$

and simplifying the square root gives

$$\frac{(\bar{S}_J + \gamma^2 b) \pm \gamma\sqrt{(\bar{S}_J+b)^2 + b^2(1-\gamma^2)(1+\lambda_J)}}{1-\gamma^2} \quad .$$

To get computable confidence limits, b is replaced by the estimate $\hat{b}$. However, this adds extra variability to the limits and tends to make them too large. To adjust and simplify, we replace the factor $1 + \lambda_J$ by 1, arriving at the final form

$$\underline{U},\overline{U} = \frac{\bar{S}_J + \gamma^2 \hat{b} \pm \sqrt{(\bar{S}_J+\hat{b})^2 + \hat{b}^2(1-\gamma^2)}}{1-\gamma^2} \quad .$$

4. EXPONENTIAL TAIL ESTIMATES APPLIED TO STATIONARY SEQUENCES

A simulation experiment was carried out to see how sensitive exponential tail fitting was to the presence of correlation. Recall that if $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ are the order statistics, then the exponential tail fit for $x \geq X_{(m)}$ has the form

$$1 - F(x) = 1 - F(X_{(m)})\, e^{-[x-X_{(m)}]/a} \quad .$$

The parameter a is estimated by

$$\hat{a} = \frac{1}{m-1} \sum_{1}^{m-1} [X_{(k)} - X_{(m)}] \quad . \tag{4.0}$$

If the exponential tail fit holds exactly, then

$$X_{(k)} - X_{(m)} = a\, E_{(k)} \quad , \quad k = 1,\ldots,m-1$$

where $E_{(1)} \geq \cdots \geq E_{(m-1)}$ have the distribution of order statistics in a sample of size m-1 from an exponential distribution with unit mean. Therefore,

$$E(X_{(1)} - X_{(m)}) = a\, E\,(E_{(1)}) = a\, u_1 \quad ,$$

where

$$u_1 = \sum_{1}^{m-1} 1/k \simeq \log\,(m-1) + \gamma \quad ,$$

where γ is Euler's constant. Thus

$$E\, X_{(1)} = E\, X_{(m)} + a\, \mu_1 \quad .$$

Estimating $E\, X_{(m)}$ by $X_{(m)}$ and a by the estimation $\hat{a}$ of Eq. (4.0) gives the exponential tail estimate

$$X_{(m)} + \mu_1 \hat{a}$$

for the expected maximum of n observations.

The stationary time series were generated as follows: Let

$$Y_{n+1} = \rho\, Y_n + \sqrt{1-\rho^2}\, e_n \ , \ n = 1,\ldots,199 \qquad (4.1)$$

$$Y_1 = e_1$$

where e_n are independent N(0,1) variables. Thus, the $Y_1,\ldots,Y_{200}$ form a Gaussian Markov chain with auto correlation

$$E\, Y_{n+k}\, Y_k = \rho^{|k|} \quad .$$

They have mean zero and variance one. The actual sequence used consisted of the lognormal variables

$$X_n = e^{\mu_j + \sigma_j Y_n} \ , \ n = 1,\ldots,200$$

where, as in our previous work,

$$\mu_j = -(\log 2)/2b_j \quad \sigma_j = \sqrt{\log 2}/b_j$$

and b_j took the values .5, .75, 1.0, 1.25, 1.5. This range included a very heavy tailed distribution (b = .5) and ranged up to the light tailed distributions at b = 1.25, 1.5.

Each run was repeated 1000 times using the Massaglia "Super-Duper" uniform random number generator and the Box-Mullen transformation. Since the exact value of $E\,X_{(1)}$ was difficult to compute (except for $\rho = 0$) the average value of $X_{(1)}$ over the 1000 runs was taken as the target figure. Examination of the SD's of $X_{(1)}$ showed that the error in using $\bar{X}_{(1)}$ as an estimate of $EX_{(1)}$ would not appreciably affect the results.

As in our other work, the $SD[X_{(1)}]$ was taken as the benchmark. The RMS error at the exponential tail estimate using m = 30 was computed using the set of ρ values

$$\rho = 0, \pm.2, \pm.4, \pm.8 \quad ,$$

and divided by $SD[X_{(1)}]$ to give a measure of the improvement in using the exponential tail estimate instead of $X_{(1)n}$ as an estimate of $EX_{(1)}$.

First of all, it is interesting to note how $EX_{(1)}$ [or rather $\bar{X}_{(1)}$] is affected by the correlation. This is tabled below:

$\bar{X}_1$

	b				
	.5	.75	1.0	1.25	1.5
.8	48.6	12.1	6.3	4.3	3.3
.4	60.9	14.4	7.2	4.8	3.7
ɔ .2	61.3	14.6	7.3	4.9	3.7
.0	61.3	14.6	7.3	4.9	3.7
-.2	61.4	14.6	7.3	4.9	3.7
-.4	61.4	14.6	7.3	4.9	3.7
-.8	60.9	14.3	7.1	4.8	3.6

The actual $\bar{X}_1$ values are quite insensitive to ɔ except at ɔ = .8.

RMS(EST)/SD[$X_{(1)}$]

	b				
	.5	.75	1.0	1.25	1.5
.8	.53	.66	.74	.80	.86
.4	.49	.58	.58	.57	.57
p .2	.52	.60	.58	.56	.55
.0	.55	.61	.58	.55	.54
-.2	.54	.60	.58	.55	.54
-.4	.55	.60	.57	.55	.54
-.8	.47	.56	.58	.60	.62

Interestingly enough, the largest effect of correlation is on the light-tailed distributions, most pronouncedly for large positive correlation. On examining the statistics, the reason for the loss of efficiency is not that the bias has increased. In fact, the exponential tail estimates have very little bias at $p = \pm .8$ for the light-tailed distributions. The problem is that their variances increase considerably.

At any rate, the exponential tail estimates hold up fairly well, always have an RMS error less than that of $X_{(1)}$, and are a considerable improvement uniformly for the heavy tailed distributions. Of course, exponentially decreasing auto-correlation in our example rules out long term dependence, and one could manufacture examples of stationary sequences with long term dependence where the exponential tail estimate would give very poor results.

REFERENCES

1. Breiman, L., J. Gins and C. Stone (1978), "Statistical Analysis and Interpretation of Peak Air Pollution Measurements," TSC-PD-A190-10, Final Report on U. S. Environmental Protection Agency Contrac. No. 68-02-2857, Technology Service Corporation, Santa Monica, California.

2. Breiman, L., C. J. Stone and J. P. Gins (1979), "New Methods for Estimating Tail Probabilities and Extreme Value Distributions," TSC-PD-A226-1, Technology Service Corporation, Santa Monica, California.

3. DuMouchel, W. H. and R. A. Olshen (1975), "On the Distribution of Claims Costs," in Credibility, Academic Press, New York, pp. 23-50.

Appendix I

MEAN AND VARIANCE CALCULATIONS

This appendix contains the calculations leading to the mean and variance, under the quadratic model, of the statistic

$$\hat{\theta} = \sum_{1}^{m-1} k\omega_k [X_{(k)} - X_{(k+1)}] \quad .$$

Put $W_0 = 0$ and $\gamma_k = k\omega_k - (k-1)\omega_{k-1}$ to get

$$\hat{\theta} = \sum_{1}^{m-1} \gamma_k L(E_{(k)})$$

$$= a \sum_{1}^{m-1} \gamma_k E_{(k)} + \frac{b}{2} \sum_{1}^{m-1} \gamma_k E_{(k)}^2 \quad .$$

We use repeatedly the fact that the $E_{(k)}$ have the representation

$$E_{(k)} = \sum_{j=k}^{m-1} \frac{Z_j}{j}$$

where $Z_1, \ldots, Z_{m-1}$ are independent exponential variables with mean one. Therefore

$$E(E_{(k)}) = \mu_k$$

and

$$E(E_k^2) = E\sum_{j,\ell=k}^{m-1} \frac{Z_k Z_\ell}{j\ell}$$

$$= \sum_{j,\ell=k}^{m-1} \frac{1}{j\ell} + \sum_{j=k}^{m-1} \frac{1}{j^2}$$

$$= \mu_k^2 + \mu_k^{(2)} \quad .$$

Hence

$$E\hat{\theta} = a \sum_1^{m-1} \gamma_k \mu_k + \frac{b}{2} \sum_1^{m-1} \gamma_k [\mu_k^2 + \mu_k^{(2)}] \quad .$$

For any sequence β_k, $k = 1, \ldots, m-1$

$$\sum_1^{m-1} \beta_k \gamma_k = \sum_1^{m-1} k\omega_k(\beta_k - \beta_{k+1}) \qquad (I.1)$$

using the convention $\beta_m = 0$. Hence

$$\sum_1^{m-1} \gamma_k \mu_k = \sum_1^{m-1} \omega_k$$

and

$$\sum_{1}^{m-1} \gamma_k [\mu_k^2 + \mu_k^{(2)}] = \sum_{1}^{m-1} k\omega_k (\mu_k^2 - \mu_{k+1}^2 + \frac{1}{k^2})$$

$$= 2 \sum_{1}^{m-1} \mu_k \omega_k \quad ;$$

so

$$E\hat{\theta} = a \sum_{1}^{m-1} \omega_k + b \sum_{1}^{m-1} \mu_k \omega_k \quad .$$

To get the variance of $\hat{\theta}$, write

$$\hat{\theta} = a \sum_{1}^{m-1} \omega_k Z_k + \frac{b}{2} \sum_{k=1}^{m-1} \gamma_k \left(\sum_{j,\ell=k}^{m-1} \frac{Z_j Z_\ell}{j\ell} \right)$$

and denoting

$$h_{j\ell} = \frac{1}{2} \sum_{1}^{\min(j,\ell)} \frac{\gamma_k}{j\ell}$$

gives

$$\hat{\theta} = a \sum_{k=1}^{m-1} \omega_k Z_k + b \sum_{j,\ell=1}^{m-1} h_{j\ell} Z_j Z_\ell \quad .$$

Let $\tilde{Z}_k = Z_k - 1$. Then

$$\hat{\theta} - E\hat{\theta} = a \sum \omega_k \tilde{Z}_k + 2b \sum_{j,\ell} h_{j\ell} \tilde{Z}_\ell + b \sum_{j,\ell} h_{j\ell} (\tilde{Z}_\ell \tilde{Z}_j - \delta_{j\ell})$$

where $\delta_{j\ell} = 1$ if $j = \ell$; 0 otherwise. Put $t_k = a\omega_k + 2b \sum_j h_{j\ell}$, then,

$$(\hat{\theta} - E\hat{\theta})^2 = \sum_{k,j} t_k t_j \tilde{Z}_k \tilde{Z}_j + 2b \sum_{j,\ell,k} t_k h_{j\ell} (\tilde{Z}_\ell \tilde{Z}_j - \delta_{j\ell}) \tilde{Z}_k$$

$$+ b^2 \sum_{j,\ell,k,i} h_{j\ell} h_{ik} (\tilde{Z}_\ell \tilde{Z}_j - \delta_{j\ell})(\tilde{Z}_i \tilde{Z}_k - \delta_{ik}) \quad . \quad (I.2)$$

Taking expectations and using

$$E\tilde{Z}_k \tilde{Z}_j = \delta_{kj}, \quad E\tilde{Z}_k \tilde{Z}_j \tilde{Z}_\ell = 2\delta_{kj}\delta_{j\ell}$$

gives

$$\mathrm{Var}(\hat{\theta}) = \sum t_k^2 + 4b \sum t_k h_{kk} + b^2 \sum h_{kk}^2 E(\tilde{Z}_k^2 - 1)^2$$

$$+ b^2 \sum_{\ell \neq k} h_{\ell\ell} h_{kk} \left\{E(\tilde{Z}_k^2 - 1)\right\}^2 + 2b^2 \sum_{\ell \neq k} h_{\ell k}^2 \quad .$$

A computation gives $E(\tilde{Z}_k^2 - 1)^2 = 8$, $E(\tilde{Z}_k^2 - 1) = 0$; so

$$\mathrm{Var}(\hat{\theta}) = \sum t_k^2 + 4b \sum t_k h_{kk} + b^2 \left(8 \sum h_{kk}^2 + 2 \sum_{\ell \neq k} h_{\ell k}^2 \right) \quad .$$

Now,

$$h_{j\ell} = \frac{1}{2} \frac{\min(j,\ell)}{j\ell} \omega_{\min(j,\ell)} \quad ,$$

so

$$h_{kk} = \frac{1}{2} \frac{\omega_k}{k}$$

and

$$\sum_{j=1}^{m-1} h_{j\ell} = \sum_{j=1}^{\ell} h_{j\ell} + \sum_{j=\ell+1}^{m-1} h_{j\ell}$$

$$= \frac{1}{2} \frac{1}{\ell} \sum_{1}^{\ell} \omega_j + \frac{1}{2} \omega_\ell \mu_{\ell+1}$$

$$= \frac{1}{2} \gamma_\ell + \frac{1}{2} \omega_\ell \mu_{\ell+1} \quad .$$

Also

$$2 \sum_{\ell \neq k} h_{\ell k}^2 = 4 \sum_{\ell < k} h_{\ell k}^2 = 4 \sum_{\ell \leq k} h_{\ell k}^2 - 4 \sum h_{kk}^2 \quad ;$$

so denoting

$$h_\ell = \sum_{j=1}^{m-1} h_{j\ell} \quad ,$$

we have

$$\mathrm{Var}(\hat{\theta}) = \sum (a\omega_k + 2bh_k)^2 + 4b \sum (a\omega_k + 2bh_k)h_{kk} + 4b^2\left(\sum h_{kk}^2 + \sum_{\ell<k} h_{\ell k}^2\right)$$

$$= \sum [a\omega_k + 2b(h_k + h_{kk})]^2 + 4b^2 \sum_{\ell<k} h_{\ell k}^2 \quad .$$

Note that

$$h_k + h_{kk} = \frac{1}{2}\gamma_k + \frac{1}{2}\omega_k\mu_k \quad .$$

To compute:

$$4 \sum_{\ell<k} h_{\ell k}^2 = \sum_{\ell<k} \frac{\omega_\ell^2}{k^2}$$

$$= \sum_{\ell=1}^{m-1} \mu_\ell^{(2)} \omega_\ell^2$$

where

$$\mu_\ell^{(2)} = \sum_{k=\ell}^{m-1} \frac{1}{k^2} \quad .$$

So finally we have the equation for the variance given in the text:

$$\mathrm{Var}(\hat{\theta}) = \sum_{k=1}^{m-1} (a\omega_k + b\gamma_k + b\omega_k\mu_k)^2 + \sum_{k=1}^{m-1} \mu_k^{(2)}\omega_k^2 \quad .$$

AD-A095 721 TECHNOLOGY SERVICE CORP SANTA MONICA CA F/G 12/1
FURTHER DEVELOPMENT OF NEW METHODS FOR ESTIMATING TAIL PROBABIL--ETC(U)
JAN 81 L BREIMAN, C J STONE, J D GINS F49620-80-C-0037
UNCLASSIFIED TSC-PD-A243-1 AFOSR-TR-81-0136 NL

END
DATE FILMED
3-81
DTIC

Appendix II

DERIVATION OF MINIMUM VARIANCE UNBIASED ESTIMATORS AND MINIMUM SQUARED ERROR ESTIMATORS

In this appendix the minimum variance unbiased estimators and the minimum squared error estimators are derived. Writing the variance as

$$\frac{\mathrm{Var}(\hat{\theta})}{a^2} = \sum_{1}^{m-1} [\omega_k + h(\gamma_k + \omega_k \mu_k)]^2 + h^2 \sum_{1}^{m-1} \mu_k^{(2)} \omega_k^2 \quad . \qquad \text{(II.1)}$$

The conditions for unbiasedness, if

$$\theta = La + Mb$$

are

$$\sum \omega_k = L, \qquad \sum \omega_k \mu_k = M \quad . \qquad \text{(II.2)}$$

The minimum variance unbiased estimators, minimized at h, are gotten by selecting the ω_k to minimize Eq. (II.1) for a given value of h under the constraints of Eq. (II.2).

For any set of coefficients ω_k, the bias at h is

$$B = a\left(L + hM - \sum \omega_k - h \sum \omega_k \mu_k\right) \quad .$$

To get the minimum squared error estimate at h, minimize

$$\frac{[\mathrm{Var}(\hat{\theta}) + B^2]}{a^2} \quad . \qquad \text{(II.3)}$$

An exact solution seems formidable. We get an "almost" solution by noting that the second term in Eq. (II.1) is usually small in size compared with the first (except, perhaps, for negative h). Hence, we throw it away and work on minimizing

$$\sum_{1}^{m-1} [\omega_k + h(\gamma_k + \omega_k \mu_k)]^2$$

with the constraints of Eq. (II.2) or with B^2/a^2 added.

Let

$$Z_k = \omega_k + h(\gamma_k + \omega_k \mu_k) \quad ;$$

then

$$\sum Z_k = \sum \omega_k + h \sum \gamma_k + h \sum \omega_k \mu_k \quad .$$

Verifying that

$$\sum_{1}^{m-1} \gamma_k = \sum_{1}^{m-1} \frac{1}{k} \sum_{1}^{k} \omega_\ell = \sum \omega_k \mu_k$$

gives

$$\sum Z_k = \sum \omega_k + 2h \sum \omega_k \mu_k \quad .$$

Define the sequence α_k, k = 1, ..., m-1 to satisfy the identity

$$\sum_{1}^{m-1} \alpha_k Z_k = \sum_{1}^{m-1} \omega_k \mu_k \quad .$$

Then

$$\sum \omega_k = \sum Z_k - 2h \sum Z_k \alpha_k \quad .$$

The minimum variance unbiased problem reduces to Eq. (II.4):

$$\text{Minimize } \sum Z_k^2 \text{ subject to}$$

$$\sum Z_k = L + 2hM$$

$$\sum \alpha_k Z_k = M \quad . \tag{II.4}$$

The minimum squared error problem becomes Eq. (II.5):

$$\text{Minimize } \left(\sum Z_k^2 + \tilde{B}^2 \right)$$

where

$$\tilde{B} = \left(L + hM - \sum Z_k + h \sum \alpha_k Z_k \right)^2 \quad . \qquad \text{(II.5)}$$

In the first problem, the solution is clear

$$Z_k = \lambda_0 + \lambda_1 \alpha_k$$

where λ_0, λ_1 are determined by

$$(m - 1)\lambda_0 + \lambda_1 \sum \alpha_k = L + 2hM$$

$$\lambda_0 \sum \alpha_k + \lambda_1 \sum \alpha_k^2 = M \quad . \qquad \text{(II.6)}$$

In the second problem, we get the minimizing equation

$$Z_k = \tilde{B}(1 - h\alpha_k) \quad ;$$

so the solution has the same form as the solution to the first problem. We can solve for $\tilde{B}$ by using

$$\sum Z_k = \tilde{B}\left(m - 1 - h \sum \alpha_k \right)$$

$$\sum \alpha_k Z_k = \tilde{B}\left(\sum \alpha_k - h \sum \alpha_k^2 \right) \quad ;$$

so

$$\tilde{B} = L + hM - \sum Z_k + h \sum \alpha_k Z_k$$

$$= L + hM - \tilde{B}\left(m - 1 - h\sum \alpha_k\right) + h\tilde{B}\left(\sum \alpha_k - h\sum \alpha_k^2\right) \quad ,$$

or

$$\tilde{B} = \frac{L + hM}{m - 2h \sum \alpha_k + h^2 \sum \alpha_k^2}$$

$$= \frac{L + hM}{1 + \sum (1 - h\alpha_k)^2} \quad .$$

In both cases the solution hinges on solving for α_k, and then being able to use the Z_k values to get the ω_k values.

Write $s_\ell = \sum_1^\ell \omega_k$ so

$$\omega_k = \Delta s_k = s_k - s_{k-1};\ s_0 = 0 \quad .$$

Then

$$Z_k = (1 + h\mu_k)\Delta s_k + \frac{h s_k}{k}$$

and the identity $\sum \alpha_k Z_k = \sum \omega_k \mu_k$ can be written as

$$\sum \alpha_k \left[(1 + h\mu_k)\Delta s_k + \frac{h s_k}{k} \right] = \sum \frac{s_k}{k} \quad .$$

Write, assuming $\alpha_m = 0$,

$$\sum_{1}^{m-1} \alpha_k \mu_k \Delta s_k = \sum_{1}^{m-1} s_k (\alpha_k \mu_k - \alpha_{k+1} \mu_{k+1}) \quad ;$$

define $\Delta_+ \beta_k = \beta_k - \beta_{k+1}$, so getting the equation

$$\sum s_k \left[\Delta_+ [(1 + h\mu_k)\alpha_k] + \frac{h\alpha_k}{k} \right] = \sum \frac{sk}{k} \quad .$$

The identity will follow if

$$\Delta_+[(1 + h\mu_k)\alpha_k] + \frac{h\alpha_k}{k} = \frac{1}{k}, \; k = 1, \ldots, m-1$$

or

$$K\Delta_+[(1 + h\mu_k)\alpha_k] + h\alpha_k = 1, \; k = 1, \ldots, m-1 \quad . \qquad \text{(II.7)}$$

Denote $q_k = 1 + h\mu_k$ and look at the homogeneous equation ($\beta_m = 0$)

$$K\Delta_+(q_k \beta_k) + h\beta_k = 0, \; k = 1, \ldots, m-1 \quad .$$

The solution β_k to this equation has two uses: First, write

$$\sum_{1}^{\ell-1} \beta_k Z_k = \sum_{1}^{\ell-1} s_\ell \left[\Delta_+ (q_k \beta_k) + \frac{h\beta_k}{k} \right] + \beta_\ell s_\ell (q_\ell + \frac{h}{k})$$

or

$$s_\ell = \frac{1}{\beta_\ell (q_\ell + \frac{h}{\ell})} \sum_{1}^{\ell} \beta_k Z_k \quad . \tag{II.8}$$

Since $\omega_k + \Delta s_k$, Eq. (II.8) gives ω_k in terms of Z_k. The other use of the β_k sequence is this: Let $\alpha_k = \sigma_k \beta_k$, then note that for any two sequences X_k, Y_k

$$\Delta_+ X_k Y_k = X_k Y_k - X_{k+1} Y_{k+1}$$

$$= X_k \Delta_+ Y_k + Y_{k+1} \Delta_+ X_k \quad .$$

Thus

$$k\Delta_+ (q_k \alpha_k \beta_k) + h\alpha_k \beta_k$$

$$= k\sigma_k \Delta_+ q_k \beta_k + k q_{k+1} \beta_{k+1} \Delta_+ \sigma_k + h\sigma_k \beta_k$$

$$= k q_{k+1} \beta_{k+1} \Delta_+ \alpha_k \quad .$$

To solve Eq. (II.7), then, we need to solve

$$k\, q_{k+1}\, \beta_{k+1}\, \Delta_+ \sigma_k = 1$$

or

$$\Delta_+ \sigma_k = \frac{1}{k q_{k+1}\, \beta_{k+1}} \quad .$$

To solve for β_k, write

$$\Delta_+ (q_k \beta_k) + \frac{h\beta_k}{k} = 0$$

or

$$(q_k + \frac{h}{k})\beta_k = q_{k+1}\, \beta_{k+1}$$

or for q k > 1,

$$\beta_k = \beta_1 \frac{\prod_1^{k-1} (q_j + \frac{h}{j})}{\prod_2^k q_j} \quad .$$

Define Π_k, k = 0, ..., m-1 by $\Pi_0 = 1$

$$\Pi_k = \prod_1^k (1 + \frac{h}{j q_j}) \quad .$$

Since β_1 is arbitrary, use the solution

$$\beta_k = \frac{\Pi_{k-1}}{q_k} \quad .$$

Then

$$\beta_\ell(q_\ell + \frac{h}{\ell}) = \Pi_{\ell-1}\ (\frac{1 + h}{\ell q_\ell}) = \Pi_\ell \quad ,$$

so Eq. (II.8) becomes

$$s_\ell = \frac{1}{\Pi_\ell}\sum_{1}^{\ell}\Pi_{k-1}\ \frac{Z_k}{q_k} \quad .$$

Going back to the σ_k sequence, we have

$$\Delta_+\sigma_k = \frac{1}{k\,\beta_{k+1}\ q_{k+1}} = \frac{1}{k\Pi_k}$$

leading to the solution

$$\sigma_\ell = \sum_{\ell}^{m-1}\frac{1}{k\Pi_k}$$

and finally, to,

$$\alpha_k = \frac{1}{k\Pi_k}\sum_{k}^{m-1}\ \frac{1}{j\Pi_j}$$

thus completing the work.

The ω_k given by the above are not easy to compute by hand; a short computer program was written to calculate them. The question then came up of whether the simple approximate solution given in Subsection 3.2.

was close enough to the complicated solution given above so that it could be used with almost equal effectiveness.

The first criterion we computed was average percentage of difference; i.e., if ω_k is the above solution at h and ω_k' is the approximate solution, define

$$\text{Avg. \% Diff.} = 100 \times \frac{\sum_{1}^{59} |\omega_k - \omega_k'|}{\sum_{1}^{59} |\omega_k|}$$

This is tabulated below.

Average Percentage of Difference									
h =	0.7	0.6	0.5	0.4	0.3	0.2	0.1	.0	-0.1
% Diff.	14.6	10.9	10.5	9.9	8.9	7.4	4.9	.00	-14.6

The approximation and the solution given above diverge for negative h. This may be due to the fact that the second term of $\text{Var}(\hat{\theta})$, which we discarded, becomes significant for negative h.

To gauge the effects of this difference we ran both solutions at h = 0.5 and h = 0.25 as estimators of $X_{(1)}$. The comparison is given below.

Comparison of Ratios

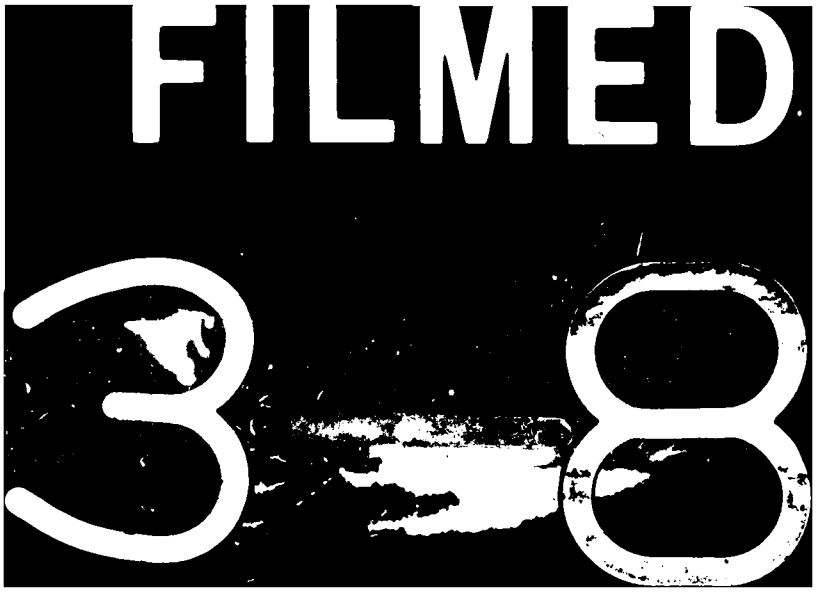
b =	.5	.75	1.0	1.25	1.50	1.75	2.0	
Exact .5	.46	1.12	2.47	3.69	4.73	5.60	6.35	Weibull
Approx .5	.44	1.09	2.33	3.46	4.42	5.21	5.90	Weibull
Exact .25	.64	.56	1.39	2.23	2.96	3.56	4.10	Weibull
Approx .25	.61	.56	1.34	2.12	2.80	3.36	3.85	Weibull
Exact .5	.54	.37	.42	.88	1.37	1.81	2.21	Lognormal
Approx .5	.52	.36	.41	.84	1.30	1.71	2.07	Lognormal
Exact .25	.59	.52	.35	.43	.70	.98	1.24	Lognormal
Approx .25	.56	.50	.34	.43	.68	.94	1.19	Lognormal

These are close, and the biases are equally close. Thus, for long-tailed distributions, which is the only solution in which the linear model produces effective estimators, the approximate solution probably can be substituted for the exact solution without a y deterioration in performance.

Comparison of Ratios

b =	.5	.75	1.0	1.25	1.50	1.75	2.0	
Exact .5	.46	1.12	2.47	3.69	4.73	5.60	6.35	Weibull
Approx .5	.44	1.09	2.33	3.46	4.42	5.21	5.90	Weibull
Exact .25	.64	.56	1.39	2.23	2.96	3.56	4.10	Weibull
Approx .25	.61	.56	1.34	2.12	2.80	3.36	3.85	Weibull
Exact .5	.54	.37	.42	.88	1.37	1.81	2.21	Lognormal
Approx .5	.52	.36	.41	.84	1.30	1.71	2.07	Lognormal
Exact .25	.59	.52	.35	.43	.70	.98	1.24	Lognormal
Approx .25	.56	.50	.34	.43	.68	.94	1.19	Lognormal

These are close, and the biases are equally close. Thus, for long-tailed distributions, which is the only solution in which the linear model produces effective estimators, the approximate solution probably can be substituted for the exact solution without any deterioration in performance.


FILMED
3-8